Naturalising intentionality
A teleological approach

Farias De Souza Filho, Sergio

Awarding institution:
King's College London

NATURALISING INTENTIONALITY:

A TELEOLOGICAL APPROACH

Sérgio Farias de Souza Filho

King's College London

PhD Thesis in Philosophy

London, 2018.

To my parents and Juliana.

There is probably no more abused a term in the history of philosophy

than "representation".

- John Searle

We, as cognitive psychologists, do not really understand our concepts of representation. We propose them, and talk about them, argue about them, and try to obtain evidence in support of them, but we do not understand them in any fundamental sense.

- Stephen Palmer

ABSTRACT

This thesis develops a teleological theory of mental representation to naturalise intentionality. Teleosemantics explains mental representation in terms of biological functions. The thesis addresses a number of foundational problems that threats the viability of teleosemantics. The first chapter, "The metaphysics of mental representation", develops a basic conception of mental representation that is designed to satisfy certain intuitive requirements (misrepresentation and original intentionality) and methodological requirements (explanatory power and ontological parsimony). The second chapter, "Naturalising intentionality", defends the thesis that mental representation is naturalistically reducible and in particular that it should be explained teleosemantically. After that, it addresses some of Tyler Burge's objections to reductionist naturalism in general and teleosemantics in particular. The third chapter, "The minimal conditions for intentionality: the problem of demarcation", considers the problem of demarcating the limits of intentionality and the objection that teleosemantics and other naturalist theories are too liberal. It adopts the method of reflective equilibrium to develop minimal conditions for intentionality based on mutual adjustments between intuitive and explanatory constraints. Finally, it rejects alternative proposals for demarcating intentionality in terms of causal independence or constancy mechanisms. The fourth chapter, "The minimal conditions for intentionality: the dual proposal", develops a specific solution to the problem of demarcation – the dual proposal for the minimal conditions for intentionality. The fifth chapter, "The content problem: in defence of producer-based teleosemantics", defends a producer-based version of teleosemantics and proposes solutions for functional indeterminacy problems facing teleosemantics.

ACKNOWLEDGEMENTS

This thesis is the result of almost four years of intense philosophical research as a PhD student at the Department of Philosophy at King's College London. My supervisor, David Papineau, has provided formidable engagement and attention throughout all these years, especially when I was incapable of properly working in virtue of a tension headache. Thank you so much for being such a great supervisor, David. Matthew Parrott was my second supervisor before moving to Birmingham. It was great to work with you, Matthew. Thank to Matteo Mameli and Matthew Soteriou for occasional supervisions and support. My former supervisors in Brazil are the main reason why I was able to study in London. My deepest gratitude to Fernando Raul de Assis Neto in Recife and Guido Imaguire in Rio de Janeiro. Finally, my gratitude for all those great philosophers with whom I had the opportunity to talk about my research throughout conferences in Europe, especially to those in Bielefeld.

I had a great time in London ever since I moved here in September 2014. London is now definitely my second home. I would not have had such a great experience without the support of the whole graduate community at King's. Especially, Mattia Sorgon (now in Alberta, Canada) and Nathan Oseroff have been very good friends. I would like also to thank Clare Moriarty, James Openshaw, Nathan (again), Mike Coxhead and David Jenkins for proofreading this thesis. I have learned a lot in the Mental Representation Reading Group, thank you all. Finally, my warmest gratitude to all my other London friends and to all my friends in Brazil.

I would not be able to finish this thesis without the medical treatment provided by Jeremy Nathan here in London and Maria da Penha in Brazil. My tension headache crisis was one of the toughest periods of my life but thanks to their support, I was able to overcome it.

All my love to my parents in Brazil for all their support. My warmest gratitude. I have

been very lucky to have such great parents. Thank you to all my family and sorry to my niece Luísa for having not been around in the first year of her life. Finally, my heartfelt gratitude to my partner, Juliana Brayner. Thank you for all your love. *Obrigado por esperar.*

Finally, my gratitude to the CAPES Foundation from the Brazilian government for my PhD scholarship. I would not be able to finish this thesis without its financial support.

TABLE OF CONTENTS

INTRODUCTION

Intentionality is a fundamental feature of mind. Consider mental representations like the belief that Herne Hill is in south London or some dog's memory of where the bone was buried. These are mental representations, i.e., mental states that have semantic properties. They represent the world in a certain way. But how is it possible for a given mental state to represent certain state? What is it for a state to be about another state? This is the problem of intentionality. Some philosophers claim that the problems of consciousness and intentionality are the two fundamental problems in philosophy of mind. The subject matter of this thesis is the problem of intentionality.

This thesis defends a naturalist view of intentionality. Mental representations are natural states. They are part of the natural order. However, paradigmatic natural states do not look representational at all. For instance, the atoms that constitute a given molecule or the leaves blowing in the Ruskin Park are natural states that do not represent anything. The task facing a naturalist theory of intentionality is to explain how, even so, some natural states are genuinely representational.

In this thesis, I defend a specific naturalist approach to intentionality – teleosemantics. According to it, mental representations are natural states with specific biological functions. The representational status of a given mental representation is explainable in terms of its specific biological function and it is in virtue of its biological function that it represents a given state. Teleosemantics is threaten by several foundational problems. The goal of this thesis is to defend teleosemantics from some of these.

The focus of this thesis is non-conceptual and sub-personal representational states. I assume hereafter, unless otherwise indicated, that I am dealing with simple non-conceptual representational states. For example, the frog that represents the presence of flies and the

representation of nearby predators by the vervet monkey. It is my hope that the teleological approach that I develop here one day may be developed in some way to more complex and sophisticated representational states like beliefs, desires and other conceptual representations, but carrying out this project lies outside the scope of this thesis.

Let me finish this introduction with an overview of what is to come in the following chapters. In the first chapter, I initially introduce the problem of intentionality. After that, I propose a basic conception of mental representation constituted by four requirements that a mental state should satisfy in order to be a mental representation. I propose two intuitive requirements – the misrepresentation and original intentionality – and two methodological requirements – the explanatory and ontological parsimony requirements. Finally, I reject a third methodological requirement – the radical error requirement. My strategy is to use the resulting basic conception of mental representations as a basis for the development of a more robust conception in the next chapters.

In the second chapter, I defend the viability of reductionist naturalism – the view that mental representations are reducible to natural states – and present my favoured reductionist approach – teleosemantics. First, I present and defend some motivations for reductionist naturalism. After that, I introduce teleosemantics which tries to establish a naturalist reduction of representational states in terms of biological functions. I finish the chapter with the assessment and rejection of some attacks by Tyler Burge on the orthodoxy of reductionism among contemporary naturalists.

The third chapter is dedicated to the problem of demarcation: what are the minimal conditions for a state to qualify as a mental representation? In other words, what are the limits of intentionality? The relevance of this problem for the debate on naturalist theories of mental representation is that they are often criticized for being too liberal. The objection is that teleosemantics and other naturalist theories consider certain states that are clearly not

representational as representational states. In this chapter, I propose a variation of the method of reflective equilibrium to develop and assess proposals for minimal conditions for intentionality. After that, I assess and reject two proposals for minimal conditions for intentionality, the causal independence proposal developed by Jerry Fodor and Ansgar Beckermann and the constancy mechanism proposal developed by Tyler Burge and Kim Sterelny.

The fourth chapter is dedicated to the positive stage of my investigation of the limits of intentionality. I develop the dual proposal which establishes two minimal conditions for a given state to represent a certain external feature – the success pattern and the constancy mechanism conditions. I argue that this proposal draws the genuine limits of intentionality in light of both intuitive and explanatory considerations.

The fifth and final chapter is dedicated to the content problem: in virtue of what does a mental representation represent a given state rather than another state? Teleosemantics tries to determine the representational content of a given mental representation in terms of its biological function. However, this core thesis is threatened by several problems, especially functional indeterminacy problems. My goal in this chapter is to develop a producer-based teleological theory of content to deal with these problems. I develop general defence of producer-based teleosemantics and defend this approach from the objection that it fails to keep the room open for enough misrepresentation cases. I propose a solution for two functional indeterminacy problems – the concertina and the distality problems. The latter is especially problematic for producer-based teleosemantics, while the former is equally problematic for other teleosemantic theories. Finally, I defend producer-based teleosemantics from one final objection – the source of error objection.

CHAPTER 1. THE METAPHYSICS OF MENTAL REPRESENTATION

1.1 A basic conception of mental representation

1.2 Intuitive requirements

1.3 Methodological requirements

John believes that Brazil is a big country. Anna told her mother that London is a wonderful city. A sentence in a book says that Romeo loves Juliet. These are cases of representations of the world in some way or another: John's belief represents Brazil as a big country; Anna's utterance represents London as a wonderful city; and the sentence represents Romeo as loving Juliet. Although beliefs, utterances and sentences are entirely different things, they all have representational power, i.e., they all have semantic properties. But they differ in the way in which they represent the world. So, John's belief represents the world in a different way from Anna's utterance since the former represents the state of affairs that Brazil is a big country, while the latter represents the state of affairs that London is a wonderful city. There are many kinds of representations. Some of them are mental states (beliefs, desires, perceptions, hopes, subpersonal representations, etc.), some are linguistic states (utterances, sentences, signs, etc.), while others are neither mental nor linguistic, like pictures or photographs.

Every representation has a representational content, that is, the way in which it represents the world. The representation draws a line in logical space between the states of affairs that it represents and those states of affairs that it does not represent. The content of John's belief is *Brazil is a big country*, so this mental state represents the state of affairs that Brazil is a big country, but no other state of affairs. In the case of beliefs and other indicative representations, the content is the truth conditions of the representation: John's belief is true if

and only if it is the case that Brazil is a big country and it is false if and only if it is not the case that Brazil is a big country. On the other hand, desires and other imperative representations do not have truth conditions, but satisfaction conditions. The content is the satisfaction conditions of the representation: Paul's desire for a pint of beer is satisfied if and only if he gets a pint of beer and it is not satisfied otherwise.

Representations give rise to a variety of puzzles, but the fundamental philosophical problem can be stated in a very simple way made famous by Franz Brentano. Representations have intentionality, they represent other states. But how can a state stand for another? That is, how is it possible for a state to be about another state? That is the problem of representation.[1] So, how can a state in John's mind represent that Brazil is a big country even if Brazil is very far away and John actually was never there? How can some marks in a book represent that Romeo loves Juliet, while some other marks in the same book represents other things or maybe don't represent at all? Representations can be about nonexistent or existent things, close or far away, abstract or concrete, etc. Indeed, it seems that they can be about everything. On the other hand, representations come in a variety of forms: utterances, marks on a sheet of paper, mental states, pictures, photographs, etc. Indeed, it appears that they can come in every conceivable form. That is the deep mystery of representation: how representational states of very different kinds can be about states of very different kinds? How is representation possible at all?

The focus of this thesis is on one specific kind of representation, the one that pertains to the mental reality: mental representation. Evidently, in order to investigate the nature of mental representations, the investigation of other representations are helpful and welcome, but only in an auxiliary way. This first chapter is on the metaphysics of mental representation: what is a mental representation and what constitutes it? In virtue of what is a given mental state a

[1] Brentano formulated this problem specifically on mental states (Cf. BRENTANO, 1874 [1995]), but since then this way of presenting the problem was generalized to all kinds of representational states.

mental representation? That is, what grounds the fact that a given mental state has representational powers? These are problems in the metaphysics of mental representation since they deal with what constitutes a mental representation and with what grounds the representation status of a given mental state (i.e., the property in virtue of which a given mental state is a mental representation). The investigation of the nature of the mental representation, of what constitutes it, aims to understand the deepest facts about mental representations and the role they play in grounding intentional explanations. This is a metaphysical matter.

1.1 A basic conception of mental representation

The first step in the investigation of the nature of mental representation is to specify what is meant by "mental representation". Notice that this is a term of art, it is not used in everyday life – the same goes for "mental content". Rather, the notion of mental representation is highly theoretical and varies from theory to theory, in such a way that there are different and incompatible notions of mental representation available. As John Searle has acutely remarked, "there is probably no more abused a term in the history of philosophy than 'representation'" (SEARLE, 1983, p. 11). In light of this fact, someone could propose that one should look for an umbrella term capable of covering all different notions of mental representation assumed by different theories of mental representation present in the literature. The goal of this strategy is to enable talk on mental representations that is capable of covering all notions of mental representation in order to specify what is common between all of them. Prima facie, that seems to be a good strategy, but the fact is that there is no substantial umbrella notion capable of covering all different notions of mental representation.

There is one candidate for this umbrella notion capable of covering all available notions of mental representation – the minimalist notion. According to it, a mental representation is just a mental state with semantic properties. Nothing else is required for a mental state to be a

mental representation. However, the minimalist notion is not substantial. It is a trivial fact that, for a given mental state to be a mental representation, it is a required condition that this state have semantic properties. But to say that a mental state should have semantic properties in order to be a mental representation is only a different way of saying that a mental state should represent in order to be a mental representation. It adds nothing more on the nature of mental representations. It is obvious that every mental representation should have semantic properties and since the minimalist notion requires nothing more than that for a mental state to be a mental representation, it follows that it says nothing substantial on the nature of mental representations. Thus, the minimalist notion is not a substantial notion of mental representation and one should look for a more substantial one.

It seems that the ultimate goal of this investigation would be to develop a definition of mental representation which establishes a list of necessary and sufficient conditions for a mental state to be a mental representation. Thus, a mental state would be a mental representation in virtue of the satisfaction of these conditions. However, I am quite sceptical about the prospects of developing a straight definition of mental representation that would provide these necessary and sufficient conditions. I think that any investigation which purports to establish sufficient and necessary conditions for a mental state to constitute a mental representation is doomed to fail. There are at least two reasons that support this sceptical conclusion.[2]

First, "mental representation" is used in a variety of conflicting ways such that it is hard to conceive a definition that would be able to embrace all of them. Second, there is always the possibility that the proposed definition will give rise to several intuitive counter-examples, either because it is too inclusive (i.e., it treats states that clearly are not mental representations

[2] William Ramsey, Barbara von Eckardt and others have defended a similar view, cf. RAMSEY, 2007; von ECKARDT, 2003.

as mental representations) or too exclusive (i.e., it treats states that clearly are mental representations as non-representational). Thus, the search for a strict definition of mental representation is not a promising strategy. But if so, how should we proceed?

I think that the best strategy is to start with the minimalist notion and then make it more substantial by adding intuitive and methodological requirements for a mental state with semantic properties to be a mental representation. In the end, this strategy will not deliver a definition of mental representation and not even a robust conception of it, but it will deliver a more substantial notion than the minimalist one and it will serve a different purpose. We can take the delivered conception as the basic one and then refine it by adding other requirements in order to deal with specific cases of mental representation. That is, to improve the delivered basic conception of mental representations with requirements that are required for specific kinds of mental representation, but not for other kinds. For instance, you can provide more specific requirements in order for a mental state to be a belief, a desire, a subpersonal representation, a perception, etc. Thus, the basic conception may serve later developments of robust conceptions of specific kinds of mental representation.

That said, how might one develop the basic notion? A good starting point is to look for requirements that a mental state should satisfy in order to be a mental representation on which basically everyone would agree. Evidently, it would be better to look for a requirement that literally everyone in this debate would agree on, but this is not a promising idea since if there is a lesson to be taken from the history of philosophy it is that everything is questionable under philosophical scrutiny. So, instead of looking for uncontentious requirements for a mental state to be a mental representation, my proposal is to focus on requirements for mental representations that are as uncontentious as possible. Thus, I will look for two kinds of requirements for mental representation. The intuitive ones which arise from our intuitive or commonsense view of what it is for a mental state to represent the world, and the

methodological requirements that arise from methodological considerations of scientific theories that posit mental representations in order to explain how cognitive systems work (e.g., cognitive science, neuroscience, psychology, etc.). The result of the establishment of intuitive and methodological requirements is what I will call a basic conception of mental representation.

But why can one not take this basic conception of mental representation as a definition of mental representation? After all, one can take the established conditions for a state to be a mental representation as defining what a mental representation is. The problem with such a move is that a definition of a state requires necessary and sufficient conditions for something to be that state and therefore necessitates the exclusion of any borderline case in which it is not clear whether something is a mental representation or not. However, that is not the case with the basic conception. It will not exclude every borderline case. There will be cases in which even though the state satisfies these conditions, it is not clear if it is a mental representation and, conversely, there will be cases in which even though the state does not satisfy these conditions, it is not clear that it is not a mental representation. So, the delivered basic conception is not definitional.

1.2. Intuitive requirements

I will state and defend two intuitive requirements for a mental state to be a mental representation, the misrepresentation and original intentionality requirements. But before doing that, it is necessary to solve the following problem. It is plainly plausible to claim that the positing of mental representation assumed by folk psychology should satisfy our intuitive requirements of what it is for a mental state to represent. After all, folk psychology is precisely our common sense view of how the mind works. Why should a scientific theory, that posits a mental representation to explain the behavior of a cognitive system, satisfy intuitive

requirements (i.e., intuitive requirements for a for a state to constitute a mental representation)? Is it not the case that since this is a theoretical posit by a scientific theory, it makes no difference if it is in accordance or it is not with our commonsense view of representation? All that matters is the unique set of properties that provided the relevant scientific positing – the mental representation – with the explanatory role that it plays in the theory. That seems to be the case in theories of natural science, so why would be things different in the case of the positing of mental representations? For instance, it makes no difference to physics whether the positing of atom or mass is in accordance or not with our intuitive views of what atoms or masses are, all that matters are the explanatory roles that the notions of atom and mass play in physics.

The difference is that the notion of representation is pre-scientific. It already has a place in our commonsense view of the world before we start to develop scientific theories and so this non-theoretical understanding of what it is for a state to represent another state constrains the sort of things that qualifies as representational states, even in the context of a scientific theory. After all, if a state is posited as a representational state by a scientific theory and it has absolutely nothing to do with what we ordinarily call a representation, why should it be called a representation at all? If the theory still calls it a representation, then one should conclude that this is just a case of homonym – two completely different states are just being called by the same name, "representation". Hence, the theoretical notions of representation should be in some way rooted or connected to some degree with our commonsense conception of representation for them to qualify as representational states.[3] That said, let's move to what I take to be two intuitive requirements for a mental state to be a mental representation.

The misrepresentation requirement

Suppose that on a dark night you are on a farm and see an animal on the horizon and

[3] William Ramsey has also highlighted this point, cf. RAMSEY, 2007, p. 24-5.

then you believe that it is a horse. As a result, you acquire the belief that this is a horse. However, as it happens this animal is not really a horse, but a cow – you have misrepresented it as a horse because it is too far away and the surrounding environment is not properly lit. Now contrast it with the following case. Suppose that there is a great amount of smoke in the middle of a forest and since smoke means fire, surely there is fire somewhere. But how literal is "smoke means fire"? Is smoke a genuine representation of fire? No. Among other reasons, smoke is not a genuine representation of fire because it cannot misrepresent fire. Natural law supports the claim that it can never be the case that you can have smoke but no fire which originated it.[4] The lesson to be drawn from the contrast between the smoke and the belief is that for a state to represent another, it is required that it has the power to misrepresent the other state.

The first intuitive requirement for a mental state to be a mental representation is that it should be possible not only for it to represent, but also for it to misrepresent. No representation without the possibility of misrepresentation – that is the motto of philosophers of mental representation. So, if the mental state is not capable of misrepresenting, then it is not a genuine mental representation. That is the misrepresentation requirement. As expected, it turns to be a restriction for the viability of theories of mental representation: if a theory is not capable of explaining how a given mental representation can misrepresent, then this theory should be summarily ruled out. This requirement implies that one can say that a mental state that seems to represent something but that lacks misrepresentation power is not a genuine mental representation at all.

Notice that the misrepresentation requirement is not only stating that there is a distinction between representation and misrepresentation. This requirement doesn't follow directly from the fact that representations make a distinction between those conditions that it represents and those that it doesn't represent. Rather, the misrepresentation requirement states

[4] For the sake of the argument, I am assuming here that natural laws are necessary.

that it should be possible for a representational state to misrepresent. Even if there is a distinction between those conditions that a state supposedly represents and those that it doesn't, if there is no room for it to misrepresent, then it is not a genuine representation. Or, to put in another way, if it is necessary that the state will not mispresent, then it is not a representation. But what about the possibility of a cognitive system that generates so precise and accurate representational states that they will never misrepresent? Here a distinction should be made between two cases. It is plainly possible that by chance a given cognitive system always produces representational states that don't ever misrepresent. That is plainly compatible with the misrepresentation requirement. What is not compatible is that it is impossible for this system to produce supposedly representational states that misrepresent. It would not be a representation at all. It is precisely this case that the misrepresentation requirement excludes.

Finally, consider the sentence "smoke means fire" again. Why can one not hold that it literally says that smoke represents fire? Why can one not hold that this is not a *façon de parler*? Under abnormal situations, it is possible to have smoke but no fire. For instance, suppose that someone puts a certain amount of smoke inside a box, transports it to another place and then opens it in order to fool everyone around: after seeing the smoke in the air, people start to look for fire. In that case, it could be claimed that smoke misrepresents fire since there is no fire around and thus smoke satisfies the misrepresentation requirement. Then, what is problematic in claiming that smoke genuinely represents fire? I think that there are at least three problems that together constitute a strong reason to reject this move.

The first one is that in this counter-example the source of the supposed misrepresentation is not the violation of the natural law that supports the correlation between smoke and fire. Rather, the source is the transport of smoke to another place to fool everyone. In that sense, smoke would only be capable of misrepresenting in abnormal situations and so the result is that *ceteris paribus*, it is not possible for smoke to misrepresent fire. Then, if one

wants to claim that smoke means fire and yet that smoke satisfies the misrepresentation requirement, one will have to claim that smoke means fire only in abnormal situations, i.e., those situations in which there is the possibility of the violation in the correlation between smoke and fire. But does it make sense?

No. Without strong principled reasons, it is not possible to hold that smoke represents fire in abnormal situations, but not under normal situations. That would be plainly arbitrary: why would smoke represent fire in certain situations but not in others? Under *ceteris paribus* conditions it is not possible for smoke to not be correlated with fire and thus it is not possible for smoke to misrepresent fire. The abnormal situations are precisely those situations in which the *ceteris paribus* conditions are violated and so it is possible for smoke to supposedly misrepresent fire. But it makes no sense in constructing a notion of representation for those situations in which the *ceteris paribus* conditions are violated – it only makes sense in constructing it for *ceteris paribus* situations. Notice that *ceteris paribus* conditions constitute a reason for claiming that a given state is a representation in certain situations, namely, those which satisfy the *ceteris paribus* conditions. But what is at stake here is precisely the opposite of it: the thesis that smoke represents fire only under situations in which the *ceteris paribus* conditions are violated (since in those situations smoke would supposedly have misrepresentational power). Therefore, smoke does not represent fire even under abnormal situations and hence that smoke does not represent fire at all.

The second problem in claiming that smoke is a genuine representation of fire is that if one accepts smoke as a genuine representation of fire, then one is forced to accept that several other natural states are also genuine representational states – rain means cloudy sky, snow means cold weather, the height of the mercury in the thermometer means temperature, etc. In fact, you are forced to accept that every state that has a correlation supported by natural law with another state genuinely represents the latter. But this conclusion is highly counterintuitive

– just because there is a strong natural correlation between two states, can it be inferred that one state represents the other? For instance, just because there is a strong correlation between snow and cold weather, can it be inferred that snow represents cold weather? This is certainly highly implausible.

Finally, the third problem is that since there are correlations almost everywhere between states that are supported by natural law, it follows that there are representations almost everywhere. But if that is the case, then the notion of representation loses its utility and distinctiveness. On one hand, it would imply that the notion of representation is almost everywhere in physical, chemical, biological explanations, etc. Thus, the notion of representation would lose its distinctiveness because it would be almost universally applicable to physical, chemical, biological states, etc. On the other hand, why would psychology and other sciences of mind that make use of intentional explanations still appeal to the notion of representation to explain a given phenomenon? This notion would be so weak that almost every system would constitute a representational system. The loss of utility is evident. In conclusion, smoke does not genuinely represent fire because otherwise the very notion of representation would lose its utility and distinctiveness.[5]

I think that the strongest reason that some people share the intuition that smoke genuinely represents fire originates in a failure to distinguish the question of whether smoke itself represents fire from the question of whether people use the presence of smoke as medium to represent the presence of fire. These are completely different questions. The first one is about the smoke in itself as being the representational vehicle of the representation of fire, while the second question is about the mental or linguistic representation harboured by people as being the vehicle of the representation of fire. On one side, if smoke itself is a genuine representation of fire, then it represents fire no matter whether there is someone (or something) that uses it as

[5] I thank Prof. Matthew Parrott for valuable suggestions on this third problem.

a representation of fire or not. On the other side, people can use the presence of smoke as a representation of fire by appealing to the natural correlation between them – but here it is clear that what is representing fire is not the smoke in itself, but people that make use of the presence of smoke in order to represent the presence of fire. Notice that the mental or linguistic representation can represent fire via the perception of smoke simply by using the correlation between fire and smoke as a medium to represent the presence of fire. Therefore, in light of the distinction between these two questions, the intuition that could provide a basis for the judgment that smoke represents fire vanishes in the air – smoke in itself does not represent anything; rather, it is people that represent fire by making use of the natural correlation between smoke and fire.

The original intentionality requirement

The second requirement for a mental state to be a mental representation is the original or underived intentionality requirement: the state should have underived or original intentionality. It is based on the distinction, made famous by John Searle, between *original* and *derived intentionality* (SEARLE, 1992).[6] Consider the following sentences:

(1) The president believes that it will rain tomorrow.

(2) In Portuguese, "vai chover amanhã" means *it will rain tomorrow.*

Sentence (1) is used literally to ascribe a representational state to the president, namely, a belief. If this is a true sentence, then the president has a state with the representational content *it will*

[6] Sometimes underived or original intentionality is also called "intrinsic intentionality". I will avoid using this term because it may be misleading in the sense of being interpreted as implying semantic internalism: to have mental representations depends solely on the intrinsic properties of the subject; the semantic properties of mental states are non-relational. However, the distinction between derived and original intentionality is completely independent of the debate between semantic internalism and externalism.

rain tomorrow. On the other hand, sentence (2) is used to ascribe a representational content to the Portuguese sentence "vai chover amanhã", namely, *it will rain tomorrow*. What is the difference between the president's belief and the Portuguese sentence? Although they have the same representational content, the belief represents that it will rain tomorrow in virtue of its own features, because it is a representational state by itself, while the Portuguese sentence represents that it will rain tomorrow not in virtue of its own features, but in virtue of something else – our ascription of this representational content to it. This sentence represents in virtue of conventions and interpretations followed by cognitive agents. If no ascription of content to the president's belief or to the Portuguese sentence had been done, then the belief would still represent that it will rain tomorrow, while the sentence would not represent anything at all. So, the belief has original intentionality because it represents what it actually represents in a non-derivative way, while the sentence has derived intentionality because its representational power is derived from something else.

The cases above illustrate not only the distinction between derived and original intentionality, but also the thesis that beliefs, desires, subpersonal representations and other mental states have original intentionality, while sentences, utterances, maps and other public states have derived intentionality. But can one infer that all public representations have derived intentionality, while only mental representations have original intentionality? Not at all. The fact is that there are several cases of public representations that represent what they do in virtue of their own features, not in virtue of the assignment of representational powers to them by cognitive agents. Let's illustrate this with the honeybee dance.

Honeybees perform a certain dance to sign the direction of the source of nectar to other honeybees. Variations in the tempo of the dance and in the axis correspond to variations in the distance and direction of the source of the nectar. The watching bees notice the performance of the dance and goes in the signaled direction to bring nectar to the hive. So, it is clear that the

honeybee dance represents the location of the source of nectar and that it is a public representation in such a way that it signals the location of nectar. But its representational power is not derived from any other representation, rather it represents what it does in virtue of its own features – there is no assignment of representational powers to it by any cognitive agent. So, this is a public representational state with original intentionality. The distinction between derived and original intentionality is widely popular, but not uncontentious. Daniel Dennett, for instance, challenges the thesis that there are states with original intentionality and thus the thesis that mental representations have original intentionality (DENNETT, 1990). However, it is not my goal here to assess Dennett's objection or any other objection to this distinction. Rather, I shall now argue that the onus of argument is on those claiming that there is only derived intentionality. Even though this will not constitute a definitive proof that this thesis is false, the conclusion will be that *prima facie* the balance leans to the acceptance of original intentionality, not to its rejection.

Suppose that there is no original intentionality, only the derived kind. So, intentional state *A* derived its intentionality from intentional state *B* which by its turn derived its intentionality from intentional state *C* and so on – this is the derivation chain. *Ex hypothesi,* it has no state with original intentionality that constitutes the source of intentionality, i.e., the source of all intentionality that passes through the chain. Pick one state in the chain – since it has derived intentionality, there will always be another state(s) from which its intentionality is derived. Furthermore, since there is no infinite number of intentional states in the natural world, it follows that this chain is finite. Thus, intentional states are connected in a myriad of ways via finite derivation chains and there are no states which constitute the sources of intentionality of these chains. But is there something problematic in this characterization of the derivation chain?

I think that there is one fundamental problem: how is it possible for a finite derivation

chain to derive its intentional properties from one intentional state to another if there is no state which is the source of its intentionality? Note that if you have an infinite derivation chain this is not so problematic, since for any element that you pick in the chain, there will always be a previous state from which its intentionality is derived. There will be a previous state from which the intentionality is derived for every intentional state in the infinite chain. However, such a move is simply not possible in a finite chain because even though there will always be a state which is the first element in the chain (no matter how long it is), even this state has derived intentionality. But from which state did this other state in the chain derive its intentionality given that it is precisely its first element?[7] The challenge to whoever holds that there is only derived intentionality consists in explaining how the intentional state which is the first element of the chain has derived intentionality given that there is no previous state in the chain (precisely because it is finite). Here two observations are necessary. The first one is that it is of no help to appeal to another finite chain to explain the derivation of intentionality of this first element since such a move would only postpone the problem to the first element of the new derivation chain: if it is also finite, how did its first element derive its intentionality? The second observation is that it is of no help either to appeal to other derivation chains to explain the derivation of the intentionality of the first element (in the sense that it was derived from several states that are the last elements in these chains) because such a move would again only postpone the problem to the first elements of these new derivation chains.

How can this problem be solved? A possible response is to claim that the derivation chain is not linear, but circular: state *A* derived its intentionality from state *B*, that derived its

[7] Here it could be complained that there is no way of determining which states are the first and last elements of the chain because it is plainly arbitrary to maintain that a given state is the first and another state is the last element but not the inverse. That is, it is arbitrary to choose one direction of the chain as the one in which it begins and therefore that it ends on the opposite direction because you could invert this claim and maintain that it begins in the opposite direction and ends in the first direction. However, this objection is flawed. There is one criterion that shows that this is not arbitrary: the last element is the only state in the chain which derives its intentionality but do not transmit it to any state. Based on this criterion, one can determine where the chain begins and therefore where it ends.

intentionality from state *C*, …, that derived its intentionality from state *N*, …., that derived its intentionality from *A*. Thus, it is not possible to tell which state is the first element in this circular chain since every element in it can be arbitrarily taken to be the first element. Moreover, no state in the chain is its source of intentionality because every state in this circular chain derives its intentionality from the previous state and transmits its intentionality to the next state.[8]

However, the same problem arises for the circular derivation chain. How is it possible for a state in the circular chain to derive its intentionality from the previous state if the transmission of its intentionality to the next state will ultimately reach the state from which this state originally derived its intentionality? After all, derivation (and transmission) of intentionality is transitive: if *A* derived its intentionality from *B* and *B* derived its intentionality from *C*, then *A* derived its intentionality from *C*. In the end, a state that derives its intentionality from a previous state in the circular chain also transmits its intentionality to this state via the transmission of its intentionality to the intermediate states between them. But how is it possible for two states in the circular chain to simultaneously derive and transmit their intentionality to each other?[9]

A possible objection to this argument is to deny the transitive character of the intentionality derivation. In order for the derivation of intentionality to not be transitive, it would be the case that if *A* derived its intentionality from *B* and *B* derived its intentionality from *C*, then there is some aspect of the intentionality of *A* that is not present in the intentionality of *C*. But how is that possible given that *B* derived the totality of its intentionality

[8] I assume here that the transmission of intentionality is just the converse of the derivation of intentionality, in the sense that if *A* derived its intentionality from *B*, then *B* transmitted its intentionality to *A*.

[9] In the formulation of this objection, I have appealed to the relations of derivation and transmission of intentionality and I have assumed that both relations are transitive. But if for whatever reason it is doubtful whether the transmission of intentionality is a transitive relation or not, it should be highlighted that this objection is plainly formulable in terms only of the relation of derivation of intentionality. The intentionality derivation is the fundamental relation here, the transmission of intentionality is secondary.

from *C* and *A* in turn derived the totality of its intentionality from *B*? Given that by hypothesis all intentionality is derived, there is no other state from which *A* derived the totality of its intentionality except from *C*. So, the derivation of intentionality is a transitive relation.

Finally, the last response to this problem would be to state that intentionality is a holistic phenomenon in the sense that no state in isolation has intentional properties because intentionality emerges out of a collection of appropriately related states. In fact, the regression of the derivation of intentionality only gives rise to the above problem because it presupposes that internationality is not holistic. But once you realize that intentionality emerges out of a collection of appropriately related states, the regression vanishes in the air. However, this response is also problematic. If one wants to hold it, one must explain how it is possible for a completely new and distinct property, intentionality, to emerge out of states that in isolation have no intentional property. Furthermore, one must explain what kind of relation should hold between theses state in order for the emergence of intentionality to be possible, what is the minimum amount of states that enables the emergence of intentionality, etc. *Prima facie* this is not an impossible task, but it constitutes a serious challenge for whoever wants to take this route to solve the above problem.

This is a hard problem for the thesis that there is only derived intentionality. But as previously observed, I am not concluding from the above argumentation that this thesis is false. After all, there is still the possibility of the development of a theory of derived intentionality committed to the thesis that there is no original intentionality but nevertheless not threatened by this hard problem. Rather, my conclusion is that the argumentation above shows that the overall balance leans to the acceptance of original intentionality precisely because that position avoids this problem. Evidently, this position has its own problem – the explanation of how a state may have original intentionality – but it is not as problematic. Thus, in light of the hard problem which threats the viability of the thesis that there is only derived intentionality, the

overall balance leans in favour of the acceptance of original intentionality, and not its rejection. *Prima facie* we should be inclined to accept the existence of original intentionality rather than rejecting it and whoever wants to deny it has a serious challenge to overcome, i.e., to solve this hard problem.

1.3. Methodological requirements

Now let's move to methodological requirements which have a very different nature from the intuitive ones. They are in force only for theories that posit mental representations in order to explain cognition – cognitive science, psychology, neuroscience, philosophical theories, etc. I will here defend two distinct but related methodological requirements – the explanatory and the ontological parsimony requirements. After that, I will assess and reject a third methodological requirement, the radical error one.

The explanatory requirement

The first methodological requirement is that for a mental state to be a mental representation, the representational status of this state should have some explanatory role. The positing of the mental representation by a cognitive theory must earn its explanatory keep. Otherwise, the state should not be recognized as a mental representation. If the positing of a given state as a mental representation plays no explanatory role whatsoever, then it makes no explanatory difference to the cognitive system which harbours it whether it is a representational state or not and thus this state shouldn't qualify as a representation. So, it is a requirement for a given mental state to be a representation that the positing of this state as representing plays an explanatory role in the theory. This is the explanatory requirement.[10] But what kind of

[10] Explanatory requirements for the positing of mental representation have been proposed by several philosophers of mental representation, cf. RAMSEY, 2007, p. 24-34; SHEA, 2007; RESCORLA, 2013.

explanatory role justifies the positing of a mental representation by a given theory?

Let's start with a case in which evidently the positing of a mental representation by a theory does not earn its explanatory role. Consider a lectern in the center of the lecture room that just remains stationary there. One can provide the following intentional explanation for why the lectern remains there: it believes that the center of the lecture room is the center of the universe and it desires above all to stay at the center of the universe. But everyone would agree that it makes no sense to provide an intentional explanation for the fact that the lectern remains in the center of the room and they would further agree that the lectern harbours no representational state. Why? Simply because the attribution of representational states to the lectern does not earn its explanatory keep. It makes no explanatory difference for the explanation of the supposed behavior of the lectern. For instance, assignments of representational states in this case play no role either in explanatory generalizations that subsume the lectern and they have no predictive power of any kind. So, we should accept the conclusion that the positing of a mental representation should be rejected.[11]

Contrast the above example with a case in which it is clear that the positing of representational states plays a crucial explanatory role that deliver generality and predictive powers: the intentional explanation of human behaviour provided by folk psychology. Suppose that Matthew goes to the kitchen and get himself a beer whenever he plays with his lovely daughter. But why does he behave in this way? There are different kinds of available explanations. A non-intentional explanation will appeal only to physical properties, states and laws in order to explain why Matthew went to kitchen and got himself a beer. That is a good explanation as it stands, but it is not good enough for our explanatory purposes. Why do we need to appeal to representational states in order to explain Matthew's behaviour in a further way?

[11] This example was originally proposed by Daniel Dennett, cf. DENETT, 1987, p. 23.

One compelling reason is that the intentional explanation of behaviour has a generality that is simply absent in the non-intentional explanation. For instance, the non-intentional explanation is not capable of explaining why Matthew gets himself a beer whenever he plays with his daughter. It can explain why in a certain period of time Matthew goes to the kitchen and gets a beer, but it cannot explain why it happens on every occasion that he plays with his daughter; why he gets a beer but not a water, etc. On the other hand, the physical explanation cannot predict that *ceteris paribus*, Matthew will drink the only beer available in the kitchen the next time that he plays with his daughter. It is precisely here that an intentional explanation is required, and the positing of a representational state earns its explanatory keep.

Now suppose that Matthew has a desire to drink a beer whenever he plays with his daughter and that he believes that the best way to do it is to go to the kitchen and get himself a beer. Therefore, because of his belief and desire, Matthew goes to the kitchen, gets a beer and drinks it. The match between his belief and desire explains why Matthew drinks a beer whenever he is playing with his daughter and predicts that he will do it in the next time that he plays with his daughter. The intentional explanation has a generality and a predictive element which is absent in the non-intentional explanation of his behaviour. At most, the non-intentional explanation will provide individual explanations of Matthew's behaviour at different periods of time, but it is not able to establish connections between the examples and provide a unifying explanation that embraces all these different events together. By contrast, the intentional explanation is provided with powers of generalization and predictions that explain why Matthew gets himself a beer on every occasion that he plays with his daughter. In fact, this pattern generalizes in folk psychology to even higher levels. If the cognitive agent desires X and believes that the best way to get X is by doing Y, then, barring other conflicting desires, that agent will do *Y*. This is the golden rule of belief-desire psychology. It predicts that we can substitute X and Y for respectively whatever desire and belief we want and provided

that they match with each other, then the agent will do Y. Such belief-desire match gives rise to the intentional explanation of the agent's behaviour. So, mental representations give rise to intentional explanations that predict and generalize behaviour and so the explanatory requirement is satisfied in folk psychology cases.

Here it should be highlighted that I am not committing myself to the stronger thesis that mental representations participate in causal relations that generate behaviour in virtue of their contents, i.e., in virtue of what they are representing. Many philosophers, following the cognitivist revolution, deny this thesis and hold that mental representations interact causally in virtue of their syntactic properties, not in virtue of their semantic properties. But this causal interaction is faithful to the contents of mental representations, notwithstanding the fact that content is causally inert. By contrast, other philosophers have claimed that mental representations not only interact in a variety of ways but participate in these causal relations in virtue of their content. This is a very controversial issue and it is not my goal here to enter into this debate. What I am claiming is the weaker thesis that the causal roles of representations merely correspond to their contents, not that representations generate behaviors in virtue of what they represent. So, if you believe that if *P* then *Q* and you also come to believe that *P*, then these two beliefs will cause you to believe *Q*, but it is contentious whether these two beliefs cause your belief *Q* in virtue of their representational contents.

But assuming the weaker thesis that the causal roles of representations merely correspond to their contents, what kind of explanatory role would the positing of mental representations have in a cognitive theory? If there is no guarantee that content is not causally inert, there is the possibility that there is a complete causal description of the behavior of the system that does not appeal to any semantic notion. In that case, what would be the explanatory role of positing mental representations? That is to say, what explanatory role would mental representations play in a theory in which there is a complete causal description of the behavior

of the cognitive system?

This is a fundamental problem that I will assess in detail only in the third and fourth chapters when I will assess the deep nature of intentional explanations of behavior in contrast with non-intentional ones. But let me highlight here just one point. Mental representations earn their explanatory keep in a way that is independent of non-intentional explanations of behavior. The explanatory role of mental representation goes further than that. Intentional explanations earn their explanatory keep because they show how the cognitive system is related to the surrounding environment, establishing how mental representations are connected with distal states of affairs in the environment (SHEA, 2013, p. 498). This is what is peculiar to intentional explanations, what distinguishes them from non-intentional explanations. Intentional explanations state how the cognitive system is connected with objects and properties of the environment with which it interacts. By contrast, the non-intentional explanation of the system may be true irrespective of the way that it is related or connected with the environment. Content assignments explain how this interaction occurs by establishing a connection between the system and the environment in which it is embedded. It is precisely in virtue of the establishment of this connection that intentional explanations have generality and predictive powers that are absent in non-intentional explanations. Thus, the golden rule of belief-desire psychology connects the agent with the environment via its beliefs and desires (an intentional explanation of Matthew's behaviour connects him with his daughter, the kitchen and the beer), but such connection is absent in non-semantic explanations of the agent's behaviour (a syntactic explanation of Matthew's behaviour establishes no connection between him and his daughter or the beer). Evidently, this feature is not exclusive of intentional explanations in folk psychology, it is also the case with intentional explanations of non-human behaviour, as well as intentional explanations that posit mental representations at the subpersonal level.

It is the establishment of connections between the cognitive system and the external

environment by intentional explanations that is the source of their generalization and predictive powers which justify the claim that the positing of mental representations earns its explanatory keep. Evidently, it is plainly possible that an intentional explanation of a given cognitive system does not earn its explanatory keep because it is not relevant to explain how the system interacts with the environment or simply because what is explanatorily relevant is completely fulfilled by the non-intentional explanation. However, in those cases in which the system-environment connection is relevant, intentional explanations earn their explanatory keep and thus the positing of mental representations satisfies the explanatory requirement.

Finally, the following question remains: to what extent does the aforementioned intentional explanation (that is justified in the case of human behaviour) generalize to subpersonal representations and non-human organisms? That is a fundamental question, but it is not my goal to assess it in this section since my goal here is to argue that generalization and predictive power justifies the positing of representational states by an intentional explanation and folk psychology is a paradigmatic case of it. To what extent the positing of mental representation earns its explanatory keep in domains outside folk psychology is a question that I will only address later. The goal of the third and fourth chapters is to assess the demarcation problem: what are the minimal conditions that a given system should satisfy in order to be an intentional system? That said, let's move to the second methodological requirement.

The ontological parsimony requirement

If two theories have the same explanatory power over the behavior of a cognitive system with the only difference that one theory posits that the system harbours a mental representation while the other doesn't posit it, then the first theory should be favoured on grounds of ontological parsimony. The same goes on for rival theories which have the same explanatory power but differ over the amount of kinds of posited mental representation – the

theory that posits less kinds of mental representation should be favoured. Generalizing, the result is that other things being equal, if theory T1 is ontologically more parsimonious in the positing of mental representations than theory T2, then T1 should be preferred to T2. That is the ontological parsimony requirement. It is a methodological requirement for the assignment of representational status to mental states, an application of the principle of ontological parsimony to the case of mental representations. Its motivation is Ockham's razor: one's ontological commitments should be driven by the principle that entities are not to be multiplied beyond necessity.

Here the following objection may be raised against the explanatory power and the ontological parsimony requirements. Neither is as valid as the intuitive requirements because they are based on methodological considerations of theories that assign representational status to a given mental state, not on the nature of the state itself. That is true, but notice that theories of mind from psychology, cognitive science, neuroscience, etc. purport to explain cognition by positing mental representations. In light of this fact, it follows that methodological considerations apply to the positing of mental representations by these theories. Accordingly, these postulations should be assessed in terms of methodological requirements. It is also important to highlight the difference between the explanatory role and the ontological parsimony requirements. On one hand, the ontological parsimony requirement dictates that parsimony is a criterion with which to assess rival theories of cognitive systems with equal explanatory power and thus to accept or reject the positing of mental representations. On the other hand, the explanatory role requirement establishes nothing about the assessment of rival theories of mind. Rather, it establishes that the positing of a mental state by a given theory should play some explanatory role that warrants such postulation. Therefore, what they have in common is that both appeal to methodological considerations, but they differ in the methodological considerations that they appeal to in order to assess the assignments of

representational statuses to mental states by theories of cognition.

The literature on mental representation is full of defences of the explanatory role requirement, in contrast with defenses of the parsimony requirement (as far as I know, there is indeed no defence of it in the literature). It strikes me that the reason for this is that philosophers who defend the explanatory role requirement usually think that once the explanatory role of positing mental representations is justified, everything is settled and the posited mental representations should be accepted in one's ontology. That is, they usually think that the only methodological requirement necessary for the assessment of theories that posit mental representation is the explanatory one. However, that is not the case. The explanatory requirement acts locally, it assesses a theory in isolation; but another requirement that acts globally is also necessary, one that crosses rival theories in order to assess them. It is here that the parsimony role requirement comes into play. Once it is settled that the positing of mental representations by a given theory is justified in light of the explanatory role requirement, it is still necessary to check whether there is a rival theory which has the same explanatory power but doesn't posit any mental representation. If that is not the case, then the mental representations posited by the first theory should be accepted. But if that is the case, then the latter theory should be preferred on behalf of ontological parsimony and thus one should get rid of the mental representations posited by the first theory. In sum, in order to accept the positing of mental representations it is necessary to assume not only a local point of view, but also a global one.

Rejecting the radical error requirement

Finally, there is the third methodological requirement, the radical error requirement. It establishes that a theory of mental representation ought to not have the consequence that we are usually radically mistaken (a) about the representational contents of our own mental states

and (b) about the contents of the mental states of others. Here is an argument in favour of the radical error requirement proposed by Dan Ryder (RYDER, 2009, p. 252-3).

His starting point is to argue that a naturalist philosopher is not committed to an a priori grasp of mental content. That is, the naturalist is not committed to the task of elucidating aprioristically what we have in mind when we talk about mental content by coming up with necessary and sufficient conditions for a mental state to have a particular content. Rather, the naturalist inclined philosopher is committed to an a posteriori grasp of mental representation which is empirically informed by the results of cognitive science and other sciences of mind. But how does such an a posteriori approach supports the radical error requirement? Consider the case of water. We have no a priori definition of water, but we should get rid of any theory which implies that the majority of things that we think are water actually are not water (e.g., the liquid in rivers, oceans, lakes, etc.). In the absence of a reason strong enough to justify such massive error, we should conclude that this theory changes the subject matter – it is not talking about water anymore. In the same vain, we should get rid of any theory which entails that the majority of mental contents that we ascribe to ours and others mental states are radically wrong. In the absence of a reason strong enough to justify such a massive error of content assignment, the conclusion is that the relevant theory is changing the subject matter – it is not talking about mental contents at all. Thus, a theory of mental representation ought not entail that we are usually radically mistaken about the contents of our own (and others) mental states.

How plausible is this requirement? First of all, we should assess its validity by distinguishing two cases. (i) the content ascriptions that we make to our own mental states; and (ii) the content ascriptions that me make to others' mental states. I think that this is a valid requirement in relation to the content ascriptions that we make to our own mental states. After all, it is hard to defend the view that we are usually radically mistaken about the contents of our own mental states, especially because it seems that we have privileged access to our own

mental states and so to their contents. However, the validity of the radical error requirement to the content of others' mental states is problematic; its validity is at most restricted to some kinds of mental states, but not to others. Let me explain why (from now on, when talking about the radical error requirement I will restrict myself to its application to the contents of the mental state of others).

The validity of the radical error requirement is restricted to some kinds of mental representation, namely, the ones that pertain to the domain of folk psychology. It loses the justification for its validity when applied to mental representations that lie outside the folk-psychology domain, like the mental representations harboured by non-human organisms. That is the case because our practices of assigning representational contents are restricted to the domain of folk psychology, when in our everyday life we assign contents to the mental states of others. However, there is no genuine practice of assignment contents in the context of cognitive science and other sciences of mind, so the practices of content assignment cannot support the validity of the radical error requirement for mental representations that inhabit domains outside folk psychology. There are at least two strong reasons for this conclusion.

The first reason is that the scientist posits that the cognitive system harbours a mental representation and assigns to it a given content in order to explain the system's behaviour. However, we cannot call it a practice of content assignment just like it happens in folk psychology. A genuine practice presupposes that the whole community or part of it agree on the standards and criteria for content assignment, but that is not the case in cognitive science and other theories in which scientists disagree on the validity of different standards and criteria for content assignment. The second reason for the conclusion that there is no practice of content assignment in the context of sciences of mind is that the scientist's content ascriptions to subpersonal representations are made in a highly theory-dependent way, in contrast with content ascriptions in the domain of folk psychology. In light of this second reason, someone

could object that it presupposes the rejection of the theory-theory view of folk psychology which claims that folk psychology is a theory of human behaviour and that mindreading is basically an exercise in theoretical reason. But the theory-theory view makes no difference to this debate. Even assuming this view, the point is that folk psychology is not a *scientific* theory in any relevant sense, while content attributions to subpersonal representations are made in the context of sciences of mind.

It should be clear, however, that by developing these objections to the universal validity of the radical error requirement, I am not committing myself to the conclusion that massive error is possible in the case of content attributions to subpersonal representations. Rather, I am just highlighting the fact that the reason which supports the validity of the radical error requirement in the case of representation in folk psychology – the practices of content assignment – is not valid for representations that lie outside this domain. Maybe there are reasons that support the applicability of this requirement to representations outside the domain of folk psychology, but they are not certainly based on practices of contents assignment (or, to defend an even weaker view, they are not based on practices of contents assignment similar to the ones that happens in the domain of folk psychology). Therefore, my conclusion is that one cannot take for granted that the radical error requirement is universally valid for the mental states of others. At most, its validity is restricted to mental representations in the domain of folk psychology. If one wants to apply this requirement to representations outside this domain, a further justification is required.

Conclusion

In this chapter, I have proposed four requirements for a mental state to be a mental representation: the intuitive requirements of misrepresentation and original intentionality and the methodological requirements of explanatory role and ontological parsimony. They

constitute a basic conception of mental representation which serves as a basis from which a robust conception of mental representation may be developed. Putting them together, the result is the following basic conception of mental representation:

> Mental Representation**:** a mental state with semantic properties such that it (I) has the power to misrepresent; (II) has original intentionality; (III) is posited by a theory to play some explanatory role; (IV) is posited in an ontologically parsimonious way.

CHAPTER 2. NATURALISING INTENTIONALITY

2.1 Intentional naturalism

2.2 Reductionist naturalism

2.3 Teleosemantics: the basic framework

2.4 The challenge of primitivist naturalism

In the first chapter, I have proposed a basic conception of mental representation constituted by four requirements that a given mental state should satisfy in order to be a mental representation – the misrepresentation, original intentionality, explanatory and ontological parsimony requirements. This basic conception is not the end of the philosophical debate on mental representations, but the beginning of it. The goal of this chapter is to defend the viability of reductionist naturalism – the view that mental representations are reducible to natural states[12] – and to present my favoured reductionist approach – teleosemantics. I will assess some foundational problems for reductionist naturalism. The first one concerns its viability, and I will establish some motivations for carrying on with it. After that, I will present teleosemantics which tries to establish a reduction of representational states to natural states in terms of the notion of biological function. Finally, I will address some objections to the orthodoxy of reductionist naturalism in contemporary philosophy of mind. There is a growing challenge to reductionist naturalism from primitivist naturalists that claim that mental representations are primitive and so irreducible natural states. I will assess and reject the attacks of Tyler Burge, the most prominent primitivist naturalist, who attacks the motivations behind reductionist

[12] Reductionist naturalism is assumed by the majority of contemporary philosophers of mental representation, cf. DRETSKE, 1981, 1988; BLOCK, 1986; FODOR, 1987b, 1990; MILLIKAN, 1984, 2004; PAPINEAU, 1984, 1993.

naturalism and the viability of teleosemantics as a reductionist theory (BURGE, 2010). That said, let's start by introducing the naturalist view of intentionality.

2.1 Intentional naturalism

The years between late 1970's and mid 1990's were the heydays of intentional naturalism in contemporary philosophy of mind. The problem of mental representation was in fashion and there was a contagious optimism that sooner or later a successful naturalist reduction of mental representation would be developed. The problem of intentionality seemed to be a fundamental philosophical problem that finally would be solved. The days of deep mystery of intentionality were about to end – philosophers would finally settle their account with Brentano.[13] However, since then things have changed. The times of inveterate naturalist optimism are gone. It is true that naturalism is still the orthodoxy among philosophers of mental representation, but the optimism that marked its heydays are gone.[14] So, what are the current prospects of the naturalist enterprise? Was it just the product of a momentary enthusiasm or is it still viable? I will defend the viability of the naturalist enterprise from the attacks of those who oppose it and argue why it is still necessary to develop a naturalistic reduction of mental content and representation. But first, it is necessary to explain what intentional naturalism is.

The first thing to notice on any investigation of philosophical naturalism is that "naturalism" has no precise meaning in contemporary philosophy. Rather, it is used in several and sometimes incompatible ways. Furthermore, depending on how the term is used, the notion of naturalism will have stricter or looser senses. In what follows, I will restrict myself to two

[13] For instance, Stephen Schiffer defended that it is an "urgent question" how the semantic and the psychological are related to the physical since "we should not be prepared to maintain that there *are* semantic or psychological facts unless we are prepared to maintain that such facts are completely determined by, are nothing over and above, physical facts" (SCHIFFER, 1982, p. 119).

[14] For instance, Jerry Fodor that once was an inveterate optimist, later adopted a more pessimistic position: "I don't want to pursue […] the question just which causal-cum-nomological relations are content-making. Those of you who have followed the literature on the metaphysics of meaning that Fred Dretske's book *Knowledge and the Flow of Information* inspired will be aware that that question is (ahem!) mootish." (FODOR, 2013, p. 12).

different and incompatible notions of naturalism.

The first one is methodological naturalism: the metaphilosophical thesis that philosophy and natural science are engaged in essentially the same practice and enterprise, pursuing similar goals and using similar methods. W.V.O. Quine is the paradigmatic methodological naturalist of contemporary philosophy, claiming that the difference between philosophy and science is not a distinction of kind, but of degree (QUINE, 1960). There is no distinctively philosophical standpoint from which philosophers can do their jobs, the philosophical and scientific enterprises are similar and often overlap. However, my focus here is not on naturalism as a metaphilosophical thesis, but with the second notion – ontological naturalism. From now on, by "naturalism" I will just mean ontological naturalism.

Ontological naturalism is the thesis that everything is natural, there is nothing outside the natural order. There is no non-natural entity, there are only natural states, objects, properties, events, etc. A complete description of reality requires only the appeal to natural things, nothing else is required. Natural properties are ontologically fundamental, there is nothing non-natural at the most fundamental level of reality. Every instantiated ontologically higher property is derived from primitive natural properties. In short, naturalism is the view that ontology – the complete inventory of reality – is exhausted by natural states and properties. Evidently, such a characterization of naturalism depends on the relevant conception of natural properties. So, what are natural properties?

There are distinct conceptions of natural properties that lead to distinct characterizations of ontological naturalism. The weakest conception is that natural properties are the ones that are not supernatural and hence that there are no supernatural entities like deities, demons or spirits. The problem with this conception is that it makes naturalism an uninformative view, after all the great majority of contemporary philosophers reject supernatural entities and so they would all count as naturalist philosophers (even though several rejects this label). A stronger

conception of naturalism is hence required. In what follows, I will present two such conceptions.

The first one is the scientific conception. According to it, natural properties are the ones which constitute the subject matter of natural sciences; i.e., the properties recognized by natural sciences. The natural order is precisely the scientific order. Supernatural properties like spirituality and divinity are not invoked by natural sciences and hence spirits and deities do not exist. By contrast, the properties of electric charge and organic matter are natural since they are invoked by natural sciences. However, such a characterization of naturalism strikes many as unsatisfactory unless it is specified what is distinctive of natural sciences – in virtue of what is a given domain of investigation a natural science? Are physics, biology and chemistry the only natural sciences or are psychology, sociology and history also natural sciences? It is not my goal to enter into this debate here, I will just highlight that the most popular view is that physics, chemistry and biology are the paradigmatic natural sciences, which shows that for the scientific conception thus conceived, natural properties are the ones recognized by these sciences.[15]

The second conception is the causal one. It claims that natural properties are the ones instantiated by entities that that have causal powers in space and time. That is, entities that constitute the causal order of causes and effects. The causal conception of natural properties identifies the natural order with the causal order and then claims that there is nothing outside the natural order thus understood. Hence, there are no numbers or sets since they are abstract entities and so lack causal powers. By contrast, properties like weight or roughness are natural since they are instantiated by entities with causal powers.[16]

Here I will remain neutral on the dispute about which of these two conceptions of natural properties is the appropriate one. The following description of intentional naturalism

[15] G.E. Moore is famous for assuming this scientific conception (even though he was not a naturalist): "By 'nature', then, I do mean and have meant that which is the subject matter of the natural sciences and also of psychology. It may be said to include all that has existed, does exist, or will exist in time." (MOORE, 1903, p. 92).
[16] Several philosophers have embraced the causal conception, cf. CRANE, 2016, p. 113.

and the motivations that I will defend for reductionist naturalism in the next section are compatible with both conceptions, so my commitment to either conception is not required.

Intentional or semantic naturalism consists in the view that semantic properties are natural properties. It is well characterized in light of the problem of intentionality: in virtue of the instantiation of which property does a given state represent another state? How is it possible for a state to be about another state? According to intentional naturalism, states represent other states in virtue of the instantiation of natural properties; in order for a state to be a representational state, it should have certain natural properties. The explanation of intentionality is purely natural, it appeals only to natural properties.[17]

Intentional naturalism is committed to realism about semantic properties – the view that there are states in the world that instantiate semantic properties. There is a semantic fact in virtue of which a sentence that assigns a semantic property to a given state has truth-conditions. For instance, the sentence "the representational state R represents *x*" is true if and only if it is a semantic fact that R represents *x*. In other words, it is true if and only if R instantiates the relevant semantic property in virtue of which R represents *x*. By contrast, intentional irrealism claims that there is nothing in the world that instantiates semantic properties and hence there are no semantic facts. Intentional irrealism strikes many as highly implausible. Words were not spared to alarm for the urgency of vindicating intentional realism. So, Jerry Fodor insisted that "if commonsense intentional psychology were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species" (FODOR, 1987b, p. xii). In the same vein, Dretske defended that the truth of intentional irrealism entails that we would have "to relinquish a conception of ourselves as human agents" (DRETSKE, 1988, p. x). However, once again, it is not my goal to enter into the debate between intentional realism and

[17] Notice that ontological naturalism implies intentional naturalism but not the other way around. It is perfectly possible to hold that semantic properties are natural while denying that everything that exists is natural (e.g., maybe mathematical entities are abstract and hence non-natural).

irrealism. Here I am just highlighting that intentional naturalism thus conceived is a realist view about intentionality, since it is committed to the theses that there are states that instantiate semantic properties and that semantic properties are natural properties.[18]

Intentional naturalism comes in two different versions depending on which level they locate intentional states in the ontological hierarchy of natural states. Primitive natural states are at the bottom level of the hierarchy and higher levels natural states at the top. According to reductionist naturalism, semantic states are higher natural states that are reducible to more primitive natural states. Hence, semantic states are fully explainable, at their most fundamental, in terms of non-semantic states and properties, with the proviso that the relevant non-semantic states and properties are natural. By contrast, primitivist naturalism maintains that semantic states are primitive natural states. There is no sense in carrying on with the reductionist enterprise of trying to reduce semantic states to more fundamental natural states because semantic states are primitive natural states. They are not reducible to any other states. Primitive natural properties are basic features of reality, they cannot be further explained by other features of reality. For instance, physics establishes that energy is primitive; it cannot be further explained by other physical notions. Primitivists claims that the same goes on with representational states. So, reductionist and primitivist naturalisms diverge on the ontological level of representational states. The former holds that they are higher natural states reducible to more primitive natural states, while the latter holds that they are not reducible to any other natural state. That is the fundamental divergence between primitivist and reductionist intentional naturalisms.

The naturalist enterprise consists in the endeavour of explaining semantic properties

[18] In principle, it is possible to coherently claim that semantic properties are natural but there are no semantic facts since nothing instantiates semantic properties. For instance, by claiming that semantic properties are uninstantiated primitive natural properties. However, as far as I know there is no philosopher of mental representation that defends this bizarre view. So, intentional naturalism as here conceived is committed to the thesis that there are states that instantiate semantic properties and thus is it is an intentional realist view.

and states in fully naturalist terms, i.e., without appealing to any non-natural notion. In its heyday, there was a quasi-consensus on the urgency of naturalizing intentionality.[19] But what are the motivations behind it? In the next section, I will present the motivations for the intentional naturalist version that I will develop in this thesis, namely, reductionist naturalism. After that, I will introduce the reductionist approach that I will develop in this thesis – teleosemantics. Finally, in the last section I will reject the motivations for adopting primitivist naturalism and also defend reductionist naturalism from the primitivist attack against the motivations for reductionism and against teleosemantics as a viable reductionist approach.

2.2 Reductionist naturalism

The core thesis of reductionist naturalism is that semantic properties are reducible to ontologically more fundamental natural properties. Representational states are fully explained in non-intentional terms. This is well illustrated by Fodor:

> "sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real it must be really something else." (FODOR, 1987b, p. 97).

The idea is that when physics finishes its ontological inventory of fundamental physical properties, we will face the following dilemma: either to show that semantic properties are reducible (or at least supervene) on fundamental physical properties or to exclude semantic

[19] Daniel Dennett, Stephen Stich and Paul Churchland were the most prominent philosophers of mental representation to oppose the reductionist program, cf. DENNETT, 1987; STICH, 1983; CHURCHLAND, P. M., 1981.

properties from our ontology, *tertium non datur.* Intentional irrealism is the high price to pay for the failure of the reductionist enterprise. Things, however, are more complicated. As Stephen Stich and Stephen Laurence have pointed out, it is problematic to claim that the failure of the program of naturalizing intentionality entails the truth of intentional irrealism (STICH & LAURENCE, 1994). Furthermore, the above motivation keeps the room open for the realist view that semantic properties are real not because they are instantiated natural properties, but because they are instantiated non-natural properties. Here I will take a different route to establish three motivations for reductionist naturalism that are also motivations against primitivist naturalism. But first it is required to explain what the reduction of one state to another state consists in.

Paradigmatic successful theoretical reductions come from natural sciences. When physics establishes that light is a certain electromagnetic radiation, it reduces light to electromagnetic radiation. That is, there is nothing more for there to be a light than for there to be an electromagnetic radiation within a certain portion of the electromagnetic spectrum. Likewise, chemistry establishes that water is the chemical substance H_2O and gold is the chemical element Au. Such scientific reductions reveal the fundamental nature of light, water and gold. They tell us what these things *really* are. They establish an identification between the reduced state and the state to which it is reduced. Hence, water is identified with H_2O, light is identified with a certain electromagnetic radiation, etc.

In scientific reductions, the reduction of X to Y explains the nature of X without appealing to any term that involves or presuppose X. For instance, the reduction of water to H_2O is done without appealing to the notion of water. To reduce one state to another state is to define it in other terms, i.e., terms that do not involve the reduced state. So, the reduction of representational states to natural states cannot appeal to any intentional notion – aboutness, representation, content, etc. That is, to reduce semantic properties is to define intentionality

appealing only to natural properties, without appealing to any intentional term.

But if semantic properties are reducible to natural properties and paradigmatic scientific reductions consist in the identification of the reduced property with another property, which natural properties are the ones identical with semantic properties? Are semantic properties identical with physical properties? The identification in the aforementioned scientific reductions is type-identification, i.e., the strict identity of the reduced properties or states with the properties or states to which they are reduced (e.g., water is H_2O). So, the hypothesis is whether there is a type-identification of semantic properties with physical properties – exactly the same properties comprise the semantic and physical. That is, whether one state – representational state – is strictly identical with another state type – physical state. Very few naturalists would actually claim that there is a type-identification between them. It is very implausible that representational states are type-identical with physical states. Let's see why.

Suppose that you are thinking about flies. Now suppose that frogs and toads also have representational states of flies (there is highly supportive scientific evidence for this from neuroethology). However, there is no physical state that is tokened in these three organisms in virtue of which they represent flies. Humans, frogs and toads have different physiological constitutions and hence physical constitutions. So, there is no single physical state that is tokened in these three organisms in virtue of which they represent flies. Rather, they token distinct physical states that constitute the representational states about flies.[20] Representational states are not type-identical with physical states because they are multiply realizable by distinct physical states. The representational state may be realized by different physical states, that is why intentional systems with different physiologies may still share a given representational

[20] Notice that it is highly implausible even that all human beings that represents flies share a single physical state that is tokened whenever they have this representational state. It is much more likely that distinct physical states are tokened throughout these cases. For instance, just think about the distinct physiological constitutions involved when a child represents flies than when an old man represents flies.

state. Thus, representational states are not type-identical with physical states.[21] Rather, it is highly plausible that they are token-identical with physical states. Semantic properties supervene on physical properties.

A-properties supervene on B-properties just in case no two things can differ with respect to A-properties without also differing with respect to B-properties. No A-difference without a B-difference. So, semantic properties supervene on physical properties in the sense that states that are exactly alike in all physical properties cannot differ with respect to semantic properties. Physical indiscernibility entails semantic indiscernibility. Representational states are not type-identical with physical states, but every token of the representational state is identical with the token of some physical state since the representational state supervenes on some physical state. So, the representational state of flies supervenes upon some physical state, no matter that this representational state is realised by different physical states of organisms with different physiologies (e.g., humans, frogs, toads, etc.). Physical states metaphysically determine representational states. There is no type-identity, only token-identity.[22]

Representational states are hence not reducible to physical states, they merely supervene upon them. But to which other natural states are they reducible? It varies from reductionist theory to reductionist theory. I don't think that there is a prima facie motivation for the identification of representational states with these or those specific natural states. But there are motivations for the identification with natural states in general, such that only the further investigation will reveal to which natural states the identification with semantic properties is viable. In what follows, I will establish three motivations for carrying on with the reductionist naturalist enterprise. They do not hang on the scientific or causal conceptions of

[21] This is the classical multiple realization objection to type-identity theories of mental states applied to the type-identification of representational states with physical states, cf. BICKLE, 2013. It was originally introduced in contemporary philosophy of mind by Hilary Putnam, cf. PUTNAM, 1967.

[22] The standard view nowadays is that there is a token-identity based on the supervenience relation between representational states and natural states. Cf. MACDONALD & PAPINEAU, 2006; SHEA, 2013.

natural properties but apply equally well to both conceptions.

The first motivation is based on the causal efficiency of representational states. It is common sense that representational states have causal powers – e.g., your belief that this is a book matched with your desire to read it causes your reading of the book; the representation of the fly by the frog causes the snapping of its tongue in the fly's direction, etc. There is a great motivation for crediting representational states with causal powers. But how is that possible? The semantic property of one state being about another state looks so *sui generis* that it is hard to see how semantic properties may have causal powers. They look very different from physical, chemical or other natural properties that clearly have causal powers. After all, the representational state may represent and misrepresent reality, represent things that no longer exist or never existed, close or very far away, microscopic or macroscopic, etc. There is nothing remotely similar to these features in standard naturalist properties. For instance, physical and chemical states have no relation with things that never existed. So, how should we explain the causal efficacy of representational states?

It is here that there is a strong motivation for reductionist naturalism. If it could be shown that semantic properties are reducible to natural properties that have causal powers, then the causal efficacy of representational states is guaranteed – there is a place for them in the causal relations of the natural order. It seems that reductionist naturalism is a promising way of saving the causal powers of representational states by showing that they are reducible to natural properties that perspicuously have causal powers. This is easy to verify in light of standard naturalist theories of mental representation like causal and teleological theories.

Causal (or informational) theories reduce representational states in terms of the conditions under which they are caused. The point of departure is that the content of a given representational state is the condition that is causally responsible for its tokening. The intentionality of a given state is constituted by the causal relations between the state and the

feature of the world that is represented.[23] Evidently, representational states are causally efficacious if semantic properties are indeed reducible to causal properties. Teleosemantics, by its turn, reduces representational states in terms of the biological function that they serve. The starting point is the specification of the representational content in terms of the condition under which the resulting behaviour succeeds in achieving a given biological goal. Once again, representational states are causally efficacious if semantic properties are indeed reducible in terms of biological success. They are causally relevant because the satisfaction of the representation's truth-conditions guarantees biological success via the production of a behaviour that achieves a given distal result. Truth-conditions are identified with biological success conditions.[24]

What about the prospects for primitivist naturalism in explaining the causal efficacy of representational states? Assuming that semantic properties are primitive natural properties, it is hard to see how they may have causal powers. In that sense, the road that leads to primitive naturalism is more difficult than the one that leads to reductionist naturalism. According to the latter, representational states have causal powers since they are reducible to natural states that are causally efficient. But what are the reasons for representational states to be causally efficient given that they are primitive natural states as defended by primitivist naturalism? Here a strong argument is required in order to show that despite the unique nature of semantic properties in virtue of which they look so different from physical properties, semantic properties are primitive natural properties that are causally efficacious.

The explanation of the causal efficacy of representational states is a strong motivation

[23] Dretske and Fodor are the main proponents of causal theories, cf. DRETSKE, 1981; FODOR, 1987, 1990.

[24] This is a simplified presentation of the success pattern explanation to illustrate how semantic properties are causally relevant according to teleosemantics. Success semantics assumes this success pattern explanation, but it is not reductionist since it just presupposes that there is given pursued result without providing a specification of it, cf. WHYTE, 1990. I will come back to this issue latter when introducing teleosemantics.

for carrying on with the reductionist enterprise.[25] The second motivation is the illuminating character of the reductionist program. If it could be shown that representational states are reducible to natural states, it would be a very illuminating result since it would reveal the ultimate metaphysical nature of representational states. The theoretical reduction illuminates entities that are still obscure and puzzling – representational states – in terms of familiar and known entities – natural states. The idea is that the naturalist reduction elucidates the obscure intentional relation of one state representing another in terms of familiar relations. But that is not the only good fruit. It also establishes connections between representational states and natural states by explaining the former in terms of the latter. If successful, the naturalist reduction would explain how representational states are connected with the wider natural order. It would be a step towards scientific integration in the sense that it would integrate cognitive theories that posit representational states in order to explain the behaviour of cognitive systems with paradigmatic natural sciences (physics, chemistry and biology), showing that there is no sort of incompatibility between them. Hence, the illuminating character of theoretical reduction is certainly a strong reason for carrying on with the reductionist program. Its success would yield very valuable fruits from both epistemic and metaphysical points of view.[26]

Finally, the third and last motivation is ontological parsimony. If it could be shown that reductionist naturalism is true, this would be highly parsimonious. After all, if a given ontological inventory assumes that intentional entities are distinct from natural entities, then it has at least two fundamental ontological categories – intentional and natural entities. It greatly contrasts with an ontological inventory that is committed to intentional naturalism and so holds

[25] This motivation assumes that semantic properties are causally efficient, but this is a contentious matter. That is the most popular view (DRETSKE, 1988; BLOCK, 1989; RESCOLAR, 2015), but several philosophers deny it (STICH, 1983). So, what is the extent in which this debate weakens this motivation? It is prima facie plausible that semantic properties are causally efficacious, this is part of our intuitive conception of representation. Thus, I take that there is a motivation for theories that explain this causal efficacy, not for theories that deny it. The onus lies on those that deny this causal efficacy, not on those that assume it.

[26] The illuminating character of the naturalist reduction of representational states has been defended as a motivation for reductionist naturalism. Cf. SHEA, 2013; PAPINEAU, 2006.

that intentional entities are natural. Finally, reductionist naturalism is ontologically more parsimonious than primitivist naturalism since it claims that semantic properties are reducible to ontologically more fundamental natural properties and hence is not committed to semantic properties as a distinctive primitive natural category. So, *ceteris paribus*, reductionist naturalist theories are ontologically more parsimonious than non-reductionist theories.[27]

Together these three motivations constitute a strong reason for carrying on with the reductionist naturalist enterprise. Its success would yield very good fruits. However, contrary to the aforementioned apocalyptic statements of Fodor and Dretske, I don't think that the failure of the reductionist program would be catastrophic or would represent the collapse of commonsense psychology and sciences of mind. Rather, it would be a pity to lose all the potential good fruits yielded by at least the partial success of the reductionist program. Let's work to reap at least some of them. In the next section, I will introduce the reductionist framework that will be developed in this thesis, namely, teleosemantics.

2.3 Teleosemantics: the basic framework

Teleology (from Greek *telos*, purpose + *logia*, study) is the explanation of phenomena by the purpose they serve. Teleological theories of mental representation, or simply teleosemantics, try to explain the nature of mental representations via the purpose or goal they serve.[28] Teleosemantics is a naturalist theory and hence such explanation should be made in naturalistic terms – it has to establish a naturalist specification of the purpose of representational states. Teleosemantics is also a reductionist theory and so it should reduce representational states to more primitive natural states in terms of the purpose they serve. But

[27] Here I am assuming that all these theories are committed to intentional realism. After all, *ceteris paribus* an intentional irrealist theory is not ontologically less parsimonious than reductionist naturalism – both may assume that there are only natural entities, but while the reductionist assumes that semantic properties are reducible to natural properties, the intentional irrealist excludes semantic properties from their ontology.

[28] Ruth Garrett Millikan and David Papineau are the main proponents of teleosemantics, cf. MILLIKAN, 1984, 2004; PAPINEAU, 1984, 1993.

what is this purpose and how does one make a naturalist specification of it?

According to teleosemantics, mental representations have biological purposes. The tokening of a representational state aims to achieve a certain biological goal. Teleosemantics claims that a representational state's biological function constitutes its biological purpose. How to exactly specify the biological functions of representational states is a contentious matter that varies from teleological theory to teleological theory. Nevertheless, they share the core thesis that representational states are reducible to more primitive natural states in terms of their biological goals and that their biological functions constitute the biological goals they serve.

The best way to introduce teleosemantics is by contrasting it with causal theories of content. The crude causal theory (CCT) claims that the content of a representational state is whatever causes its tokens. Consider COW, the mental representation of a cow. According to CCT, COW's content is *cow* since its tokens are caused by cows. However, there are situations in which the representation is caused by things that it does not represent, i.e., things that are not in its extension. For instance, there are situations in which COW is not caused by cows, but by horses. So, according to CCT, COW's content is not really *cow*, but *cow or horse* – both cows and horses are in COW's extension. The problem with this move is that it precludes misrepresentation, i.e., situations in which the representational state is false since it represents a condition that does not obtain. Thus, there is no misrepresentation because situations that are candidates for misrepresentation are always transformed into non-misrepresentation situations in virtue of their very occurrences. How may causal theories rule out causes that are constitutive of content from causes that are not constitutive because they are misrepresentations? That is, how are we to include in COW's content cows that cause its tokens and exclude non-cows that also cause its tokens? The problem of misrepresentation is how to explain the very possibility of false representations. Sophisticated causal theories try to keep the room open for misrepresentation, but it is not clear that they succeed in doing so.

It is in light of the problem of misrepresentation that teleosemantics becomes attractive. It identifies the representational state's truth-conditions with its proper functioning conditions and that keeps the room open for misrepresentation – the representation is true if and only if it performs its biological function and it is false if and only if it fails to perform its biological function. The possibility of false representations is hence guaranteed precisely because biological functions are not always performed. That is the teleosemantic strategy to solve the problem of misrepresentation. But in order to fully develop it, it is required to introduce the basic teleosemantic framework. Since the notion of biological function plays a pivotal role, let's start by introducing it.

Biological functions are assigned to biological traits in order to specify their biological goals. They may be categorized into *aetiological* and *non-aetiological functions*. Let's contrast biological traits with ordinary artefacts. Clocks have the purpose of displaying the right time, they will be proper functioning if and only if they display the right time and malfunctioning otherwise. When we say that the function of the clock is to display the right time, what we are saying is that the clock was designed for this purpose, i.e., that this is its goal. We will say that it is malfunctioning in case it fails to display the right time, i.e., because it displays the wrong time or no time at all. Note that the clock may also be used to do a lot of other things such as interior decoration, but the function of the clock *qua* clock is to display the right time. Just like artefacts, biological traits have functions. The function of the clock is to display the right time and not interior decorating; the biological function of the heart is to pump blood and not to make a thumping noise. But in virtue of what does a given trait have this or that biological function?

According to the aetiological conception of biological function, the function of a given biological trait is the effect for which it was selected. The function of the trait is determined by the history of the selection of traits of this kind. Thus, a trait has a specific function in virtue of

the fact that it was designed by some selection process to have this effect. Evidently, there are several selection processes and the selection process responsible for designs of artefacts is distinct from the selection process responsible for the designs of biological traits. The selection process that gives rise to the artefact's function is intentional – there is always a designer with the intention of designing the artefact to serve a given purpose. But the selection process that designs the biological trait and gives rise to its function is non-intentional; it is biological. There is no intentional agent behind biological selection processes. They give rise to teleonomic functions, i.e., mind-independent functions. Since the goal of teleosemantics is to establish a naturalist reduction of representational states, its proponents have to specify a natural and teleonomic selection process based on which it is possible to achieve such a reduction.[29]

But first of all, what is a selection process? I will assume here that a given process is a selection provided that it satisfies the following conditions:[30]

(1) variability in the traits possessed;

(2) selection of items with certain traits;

(3) heritability of traits selected for;

There is no selection if there is no initial variation since the same selection forces that operate in a homogeneous population will have no discriminatory effect. When there is variation, the items will be selected for some trait when this trait interacts with some specific environmental feature such that items without that trait will suffer some loss. When the trait is transmitted to the offspring of the items that initially had it, the result will be an increase in the proportion of items with this trait in the population. The conclusion is that the satisfaction of conditions (1)-(3) entails that the selected trait has the function of having the effect that caused

[29] Larry Wright was probably the first philosopher to formulate the aetiological conception of function: the function of X is Z if and only if (1) Z is a consequence (result) of Xs being there; (2) X is there because it does (results in) Z. Cf. WRIGHT, 1973.

[30] This is based on R.C. Lewontin's characterization of the selection process, cf. LEWONTIN, 1970, p. 1.

the differential reproductions of items with this trait.

There are several kinds of selection process which generate different kinds of biological function. The paradigmatic selection process appealed by teleological theories is evolutionary selection. Evolutionary selection (also called "natural selection") is an intergenerational inheritance process. The selected trait is that one which historically has had an effect which increased the survival and reproduction of the species. Thus, the selected effect is determined through the history of fitness and success of reproduction of the ancestors of the organism. More formally, the evolutionary function of a trait is thus defined:

> EVOLUTIONARY FUNCTION: some effect E is the biological function of some trait X in organism O if and only if the genotype responsible for X was selected for doing E because doing E was adaptive for O's ancestors.[31]

Accordingly, the biological function of the heart is to pump blood (not to make a thumping noise) because the effect that was adaptive for ancestral hearts was to pump blood (not to make a thumping noise). Evolutionary functions are well illustrated by the phenomenon of convergence evolution – species with different lineages independently having evolved similar traits as a result of adaption to similar environments. A notorious example is streamlining body shapes in dolphins, ichthyosaur and the great white shark. These marine hunters have very similar shapes without inheriting them from a distant common ancestor. The similar aerodynamic forms are adaptive in marine environments because they facilitate quick movements that increase predatory success. Convergence evolution shows the strong power of evolutionary selection of designing different species in very similar ways in order to achieve

[31] Cf. NEANDER, 1991; MILLIKAN, 1989a.

adaptive effects.

Evolutionary selection operates at the genetic level: it is a phylogenetic process.[32] However, there are also ontogenetic selections which operate in the development of organisms, i.e., which select traits not in the evolutionary history of the organism, but in its individual developmental history ever since the fertilization. There are differentiated reproductions of traits during the development of the organism, not only on organisms over generations. For instance, there is a differentiated reproduction of behavioural outputs via classical conditioning by a learning mechanism during the organism's own development. As a result, the organism acquires the function of producing those behavioural outputs that were favoured during the learning period. So, the aetiological conception of function is not restricted to functions constituted by phylogenetic selections, it also encompasses ontogenetic ones.

The aetiological conception is also enriched by encompassing derived biological functions (MILLIKAN, 1984, p. 17-49). A given mechanism has a direct function when it was selected to have that effect. In turn, the trait produced by this mechanism has a function that is derived from the function of this producer mechanism. A trait has a direct function when it was selected to have a given effect; a trait has a derived function when its function was derived from its producer mechanism's direct function. For instance, the chameleon's camouflage mechanism has the direct function of producing camouflage patterns that match the immediate environment. Suppose that a given chameleon for the first time camouflaged into the orange colour. This particular orange pattern has the derived function of camouflaging because it is derived from the mechanism's direct function of camouflaging, no matter that this is the very first time the mechanism produced an orange camouflage.

In this thesis, I will assume the aetiological conception of biological function. But how

[32] Here I am following the traditional view according to which only genes can be inherited through the evolutionary selection process. However, this view has been called into question in recent years by those who claim that not all evolutionary selections are at bottom genetic selections, cf. MAMELI, 2004.

does teleosemantics appeal to biological functions in order to achieve the reduction of representational states to natural states? Here it is necessary to distinguish two philosophical problems on the nature of representational states. First, *the representational status problem*: in virtue of what is a given state a representational state? That is, what is the property in virtue of which it counts as representational? Second, *the content problem*: provided that a given state is representational, in virtue of what does it represents this state but not another state? That is, what determines its representational content? To illustrate this distinction, let's consider the representational state COW again. The representational status problem asks in virtue of what is COW a representational state, i.e., which property makes it a representational state. By contrast, the content problem asks in virtue of what COW's content is *cow* but not *horse* or other non-cow, i.e., which property it instantiates that determines that *cow* is its content. Teleosemantics provides a solution to both problems, claiming that the state is representational by appealing to the fact that it has a certain biological function and that it has a specific content also by appealing to its biological function.

The various teleological theories widely differ with respect to the way in which the representational content and status are determined via the biological function of the state, but they all appeal to its biological function to determine its representational status and content.[33] I will start with how the basic teleosemantic framework solves the content problem and after that I will present its solution to the representational status problem.

Consider a representational state produced by a biological mechanism. According to teleosemantics, this mechanism is capable of representing in virtue of the fact that it was designed to produce representations. However, a mechanism that was designed to do something may fail to do it. It is always an open possibility that the mechanism fails to do what it was

[33] It should be noted that several philosophers appeal to biological function only to solve the content problem, putting aside the representational status problem. It is possible to develop a teleological solution to content problem that is not committed to the teleological solution to the representational status problem since they are distinct problems. But I will treat teleosemantics here as a theory that offers a solution to both problems.

designed to do (e.g., in virtue of an internal defect). But it is precisely this case that leaves open the possibility of false representations: biological mechanisms that were selected to produce true representations sometimes fail to do it and instead produce false representations. In the case of false representations, the mechanism failed to perform its function, i.e., to produce true representations. False representations are just representations that failed to properly represent. The feature of functions that makes them an attractive source of content is that a mechanism may not always perform its function just like a representation may not always accurately represent. That is how teleosemantics solves the problem of misrepresentation: it explains misrepresentation by appealing to the fact that the producer of mental representations doesn't always perform its function of producing true representations.

The teleosemantic core thesis is the identification of a representation's truth conditions with its proper functioning conditions. It provides a powerful distinction between a situation in which the biological mechanism is proper functioning and the situation in which it is malfunctioning. Based on this, the explanation of how it is possible for representational states to misrepresent reality becomes viable. A biological trait is proper functioning when it is doing what it was historically selected to do (i.e., it performs its function) and it is malfunctioning when it is not doing what was historically selected to do (i.e., it is failing to perform its function). The circumstances in which there is proper functioning are those in which the mechanism produces true representations and the circumstances in which there is malfunctioning are those in which the mechanism produces false representations.

That is how teleosemantics explains the possibility of misrepresentations. But how does it determine representational content? That is *the basic teleosemantic framework*:

REPRESENTATIONAL CONTENT: the biological function of the representational state determines its content. The representational state's truth-conditions are reducible to its biological success conditions.

What all teleosemantic theories have in common is the thesis that content is determined in terms of the notion of biological function. However, how to properly develop this basic teleosemantic framework is a very contentious matter, for different teleosemantic theories determine content in different and conflicting ways. They take as their starting point the thesis that the representation's truth-conditions are reducible to its biological success conditions. But then they take different roads with respect to how they fully determine content. So, to illustrate how this basic teleosemantic framework may be developed, let's see how the most representative teleosemantic approach – consumer-based teleosemantics – determines content (PAPINEAU, 1984, 1993, 2016; MILLIKAN, 1984, 2004).

The content of representational state R is the external condition C that should be the case for the behavioural output B triggered by R to perform its biological function – to succeed in achieving the effect for which the mechanism that produced B was selected. The truth of R, i.e., when C obtains, guarantees that B achieves the selected effect, while in case of the falsehood of R, i.e., when C does not obtain, B does not achieve the selected effect. So, the truth of the representation is required for the performance of the biological function. Evidently, here I am assuming that the mechanism that produced B was proper functioning, i.e., produced the appropriate behaviour to achieve the selected effect. After all, in case it does not produce the right effect, it makes no difference whether C obtains or not since the selected effect will not be achieved.[34] The representation's content is thus identified with the condition under

[34] What if the mechanism produces the inappropriate behaviour and still achieves the selected effect? That would be just a lucky coincidence. This is not a situation in accordance with the selection history in which the mechanism was selected. There the mechanism was selected to produce the appropriate behaviour to achieve this effect, that is why it was selected. Millikan calls "normal situation" the situation in which the mechanism

which the triggered behaviour performs its biological function, achieving the selected effect. That is how consumer-based teleosemantics determines content. Let's illustrate it with some real examples of the paradigmatic biological function appealed in teleosemantics, evolutionary function.

The selected effects that constitute evolutionary functions are the ones which increase the fitness and reproduction of the species. Consider a beaver that splashes its tail on the water in order to signal to its fellow beavers the presence of some predator. The tail's splash triggers the avoidance behaviour of other beavers and as a result they escape from the predator. The token of the representational state is what triggers the avoidance behaviour, but what is its content? The tail's splash is the representational state and it is produced by the visuomotor system to signal the fellow beavers about the presence of predators. This is an adaptive effect since it triggers the beavers' avoidance behaviour of escaping from predators. So, the representational content is *predator* (or *danger*) because its truth – i.e., when there is in fact a predator – guarantees the adaptive effect of signalling other beavers to escape from the predator. The fellow beavers' avoidance behaviour as a result achieves the adaptive effect of escaping from the predator. Notice that only when the tail's splash tracks the presence of the predator is the avoidance behaviour adaptive. The representation's truth-conditions are identified with its fitness-conditions. The truth of the representation guarantees the evolutionary success of the behavioural output (provided that the behaviour is the appropriate one). After all, evolutionary selection would not select the visuomotor system to produce false representations since they fail to guarantee the evolutionary success of the resulting behaviour – there would be no predator around.

The producer-consumer mechanisms distinction is useful here.[35] The representational

produces the behaviour in accordance with its selection history. She requires that the relevant situation is a "normal situation" in her specification of content to rule out lucky coincidence situations. Cf. MILLIKAN, 1984.

[35] The consumer-producer distinction was introduced by Millikan, cf. MILLIKAN, 1984.

state – the tail's splash – lies between the system that produces the representation – the beaver's visuomotor system – and the system that uses or consumes the representation – the fellow beavers that respond to the tokening of the representation by having the avoidance behaviour. The biological function of the producer system is to produce the representational state whenever a certain external condition obtains, and the tokened representational states have the derived biological function of being present whenever the external condition obtains. The function of the consumer system is to produce the appropriate behaviour, whenever the representation is tokened, in order to achieve a given adaptive effect. Such successful behaviour requires the obtaining of a given external condition. The representational content is that external condition. So, the beaver's visuomotor system has the function of producing the representation whenever there is a predator and a particular token of the representation has the derived function of being present when that particular predator is present. The function of the consumer's system – the fellow beavers – is to have an avoidance behaviour in response to the token of the representation. In the beaver's example, the consumer and producer systems are different organisms, but they may be part of the same organism. For instance, consider the famous case of the frog that represents the presence of the fly and as a result snap its tongue in the fly's direction and eats it. In this case, the frog's visual system is the producer system that tokens the representational state of the fly and the frog's motor-digestive system is the consumer system that uses the representation to catch and digest the fly.

That is how consumer-based teleosemantics determines representational content. It claims that it is the adaptive effect of the behavioural output triggered by the consumer system that determines content. The content is the external condition that should be the case for the consumer's behavioural output to be adaptive. So, the tail's splash represents the predator because the presence of the predator is required for the fellow beaver's avoidance behaviour to be adaptive. After all, if there is no predator around the avoidance behaviour would be just loss

of energy. In the same vein, the content of the frog's representational state is the external condition that should be the case for the frog's behaviour to be adaptive – *food* or *nutrients*. The digested nutrients will increase the fitness and reproduction. After all, in case the represented entity is not nutritive, the behaviour will be just loss of energy.

However, this approach is not the only way in which the basic teleosemantic framework may be developed. There are other teleological theories that reject this consumer-based criterion which gives priority to the function of the consumer system in order to determine content. Instead, they propose a producer-based criterion which gives priority to the function of the mechanism that produces the representational state in order to determine content. That is the producer-based teleosemantics (DRETSKE, 1988, 1995; NEANDER, 1995; 2017; JACOB, 1997). How to properly develop the basic framework lies the heart of the contemporary debates on teleosemantics.

What about the representational status problem? In virtue of what does a given state counts as a representational state? That is, what is the instantiated property of a given state in virtue of which it is a representation? This is what the basic teleosemantic framework establishes:

> REPRESENTATIONAL STATUS: in order for a given state to constitute a representation, it is required (1) that the behavioural output triggered by the tokening of the state has a biological function; (2) that the mechanism which produces the behavioural output uses the state as a proxy for the presence of some external condition.

The first condition requires that the mechanism which produces the behavioural output was selected to produce this behaviour whenever the representation is tokened precisely in order to achieve the selected effect. That guarantees that the behavioural output has a biological

function. The second condition requires that the consumer system should use the representational state as proxy for the presence of the external condition. It uses the state as an indication or signal that the external condition obtains. Whenever a given state satisfies these conditions, it is a full-blown representation according to the basic teleosemantics framework.

So, the tail splash of the beaver is a representational state. (1) the consumer system that produces the avoidance behaviour in response to the tokening of the tail splash state has the biological function of producing the behaviour to escape from the predator. (2) the consumer system uses the state as a proxy for the presence of the predator: the fellow beavers will move in different directions relative to where the state signals the location of the predator. That shows that the fellow beavers use the state as a proxy for the location of predators. In the same vein, the frog's state is a representational state. (1) the consumer system that produces the catching behaviour in response to the tokening of the frog's state has the biological function of catching and digesting the fly; (2) the consumer system uses the state as a proxy for the presence of the fly, the frog will snap its tongue in different directions relative to where the state signals the presence of the fly.

This basic teleosemantic framework entails that amoebas, paramecia, honeybees, vervet monkeys and other very simple organisms are intentional systems that produce full-blown representational states. Once again, this basic teleosemantic framework is developed in different ways, such that there are different teleological theories of representation that are more liberal or more restrictive in their criteria for a given system to be genuinely representational. So, certain theories consider very simple systems as representational while other theories do not count them as genuinely representational.

This is the basic teleosemantic framework's strategy for solving the content and representational status problems. Its viability is challenged by several problems and objections and as a result the basic framework is developed in several and conflicting ways. It is not

possible to assess all problems and objections in this thesis. For instance, I will not assess the famous swampman objection.[36] In this thesis, I will assess the problem of demarcation that threatens the teleosemantic solution for the representational status problem in the third and fourth chapters and I will assess the functional indeterminacy problem that threatens the teleosemantic solution for the content problem in the fifth chapter. But before all that, in the next section I will assess and reject Burge's primitivist attack against reductionist naturalism in general and also against teleosemantics as a viable reductionist approach in particular.

2.4 The challenge of primitivist naturalism

Tyler Burge defends primitivist naturalism about mental representations, according to which mental representations are primitive natural states. That is, they are primitive states which are part of the natural order but are not reducible to any other more fundamental natural state (BURGE, 2010). He claims that the reason that standard naturalist theories fail to reduce mental representation to some other natural notion – biological function (teleological theories), causal relations (causal theories), etc. – is that mental representation is a primitive, irreducible natural notion. Primitivist naturalism contrasts with both primitivist non-naturalism, which claims that mental representations are non-natural primitive states, and with reductionist naturalism, which claims that they are not primitive but are reducible to more primitive natural states. Burge's account of primitivist naturalism is the most compelling and influential defence of primitivist naturalism, particularly his arguments against the teleological enterprise of reducing representational states in terms of biological function. In this section, I will argue that despite the ingenuity of the arguments, they fail to show that representational states are primitive or that the teleosemantic account of representational states is doomed to fail.

Burge's main thesis is that there is a fundamental distinction between, on one hand,

[36] For an overview of the debate on the swampman objection, cf. NEANDER, 2012.

biological functions and norms and, on the other hand, representational functions and norms (i.e., the function and norms of accurately representing reality). This distinction arises in the context of psychological explanations which posit representational states to explain the behaviour of cognitive systems and their relations with the external environment. But how should one distinguish cognitive systems that are genuinely intentional from those to which representational states are merely imposed but which, in reality, do not represent at all? Burge claims that genuine representational states are the ones that are posited by successful explanations which are not in any way replaceable by non-representational explanations. That is, the difference between representational and non-representational systems is that the representational state figures in the successful explanations of the behaviour such that it is not replaceable by a non-representational explanation because there is an explanatory gap between them. So, the positing of representational state is required to fully explain such behaviour.

Burge focuses on psychological explanations of perceptual systems, especially visual systems. He claims that representational states constitute a distinctive kind of natural state, they play a unique explanatory role in psychological explanations which cannot be assimilated to any role played by any other natural state. In contrast, reductionists maintain that the role played by representational states in psychological explanations are assimilable to the role played by some other natural state(s). So, there is a fundamental divergence between Burge's primitivist view that representational states which figure in psychological explanations are primitive and irreducible and the reductionist view (which he calls "the deflationary tradition") according to which representational states that figure in psychological explanations are reducible to some other natural state.

In order to establish primitivist naturalism, Burge develops several objections to the reductionist view. His attack is divided in two fronts. On the first one, he attacks the motivations behind the enterprise of developing a naturalistic reduction of representational

states, claiming that the motivations go precisely on the opposite direction, in favour of primitivism. On the second front, he attacks the very possibility of a specific reductionist approach – teleosemantics. I will assess each front in turn.

The attack on the motivations for reductionist naturalism

Burge's assessment of the motivations for reductionist naturalism starts with his claim that there is a notion of representational state that is distinctive of psychological explanations which cannot be assimilated by any notion from non-psychological explanations. However, this is precisely what reductionist naturalists try to do: since they assume that representational states come in degree (from very simple systems like bacteria or amoebas to complex systems like human cognition), they tend to ignore the relevant explanatorily distinction between the role played by "representation" in psychological explanations and the role played by it in non-psychological explanations. The use of the same term in two different sorts of explanation is misleading because it overshadows the fact that the term has two completely distinct explanatory roles in the two explanations and refers to two different kinds. Burge claims that while in the case of psychological explanations the notion of representation plays an explanatory role which is distinctively psychological and genuinely representational, things are different in the case of explanations of simple organisms. There the notion of representation is dispensable and perfectly replaceable by non-representational notions (e.g., sensitivity, discrimination, covariation, registration, etc.) without at all weakening the explanatory power of the explanation. In sum, the fact that the same term is being used to refer to two different notions invoked by two different sorts of explanations is misleading because it blurs the distinction of kind between them – a representational notion in the case of psychological explanations and a non-representational notion in the case of non-psychological explanations.

But why does Burge claim that representation constitutes a distinctive state in

psychological explanations and that the motivations in favour of reducing it to more fundamental natural states are misplaced? Burge argues that reductionist naturalism is out of sync with scientific knowledge and practice (BURGE, 2010, pp. 296-7). It mistakenly assumes that a representation is not a scientifically respected notion because it is entrenched only in folk psychology and then that it must be made scientifically respectable by reducing it to some natural state which is scientifically respectable. Rather, the notion of representation is entrenched in psychological explanations and it cannot be taken to be *prima facie* defective or in need of supplementation because it has long earned its explanatory keep precisely by figuring in successful psychological explanations. According to Burge, this fact shows that the notion of representation is scientifically respectable and so that there is no urgency of reducing it to any other scientific notion. Psychological explanations are committed to entities needed to make their explanatory claims true, among which are representational states, and there is no reason to hold that a naturalist reduction of representation is required in order for psychological explanations to be successful. Their success is independent of any reductionist enterprise.

Burge argues that our ontology should be dictated by our successful scientific explanations[37] and since our successful psychological explanations posit representational states in order to explain psychological phenomena, it follows that the motivations go in the direction of accepting these representational states in our ontology. After all, Burge argues, we cannot have a better reason to rely on a notion than that it figures centrally in a successful science.[38] Successful explanations are the one which yield agreement in the scientific community, open new questions, engender improvement in theory and experimentation, achieve pragmatic results, etc. Burge appeals to mainstream theories in perceptual psychology (and visual science

[37] He develops this general claim in his paper "Mind-body causation and explanatory practice" (BURGE, 1993).
[38] "Notions like representation earn their keep in science […] by figuring in successful explanation. Successful explanation is marked in the usual ways by yielding agreement, opening new questions, making questions testable and precise, engendering progressive improvement in theory and experimentation. Mainstream work in perceptual psychology displays these features. [...] One could hardly have better epistemic ground to rely on a notion than that it figures centrally in a successful science." (BURGE, 2010, p. 298).

in particular) as a notorious example of successful psychological explanations in which representational states play a central role. So, we should accept them in our ontology.

But if representations are irreducible primitive states[39], what is left for the naturalist philosopher of mental representation to do? According to Burge, to determine the place of representations in the wider natural order by finding systematic connections between them. There is no conflict between representational states and the natural realm. The philosopher's job is to clarify, explore and connect representational states with the wider natural order.[40]

Does Burge's argument succeed in showing that the motivations behind reductionist naturalism are misplaced and that we have motivations for accepting representations as primitive natural states? No. In what follows I will try to show why Burge's objection is flawed. My response will be divided into two fronts: on the positive one, I will show that there are strong motivations for appealing to a naturalistic reduction of mental representation even assuming, like Burge, that representation is a respectable scientific notion; on the negative front, I will show that there is a fundamental failure in Burge's objection when he tries to establish that there is a motivation for adopting primitivism, not reductionism.

Following Burge, let's assume that representation is a respectable scientific notion in psychology and hence that there is no need to reduce it to a respectable scientific notion in order to make it scientifically respectable. But now suppose that there is an available successful theoretical reduction of representational states to more fundamental natural states. Since we have the prior assumption that the notion of representation is scientifically respectable

[39] "Is reduction of the sort expected by the Deflationary Tradition [i.e., reductionist naturalism] possible? Reductions are a legitimate type of explanatory unification. Occasionally reductions succeed. In principle, representation might be somehow reducible to other notions. I believe, however, that trying to reduce representation and veridicality to something more 'naturalistically acceptable' is probably pointless and hopeless. […] the notions of veridicality and representations – and notions like perceptual state, belief, propositional inference – are scientific primitives." (BURGE, 2010, p. 298).

[40] Burge also attacks apocalyptic statements of some reductionists to the effect that some dire consequence would ensue if reductionism fails (e.g., intentional irrealism). I will not assess this here since I have already previously noted that these statements are misplaced. Instead, in the second section of this chapter, I have proposed three motivations for reductionism that lack such apocalyptic feature.

regardless of whether it is reducible or not, the motivation behind this reduction is not to make the notion of representation scientifically respectable. So, is it a reason strong enough for one to discard this available reduction? No, there are at least three reasons to appeal to it.

The first reason is that the appeal to a reduction of representation is motivated not only by the need to make it scientifically respectable. Rather, it is illuminating to appeal to the reduction in order to reveal the ultimate nature of representational states. Thus, the necessity of making the notion of representation scientifically respectable is just one of the possible motivations behind the reductionist enterprise. In fact, revealing the ultimate nature of a kind of state is a motivation equally or even more strong than making it scientifically respectable.

The second reason is that Burge's own description of the philosopher of mental representation's job entails that in this hypothetical situation there is a motivation for appealing to an available theoretical reduction. Burge is right in claiming that the philosopher should determine the place of representational states in the wider natural order by finding systematic connections involving them. Progress can be made by clarifying, exploring and connecting representational states with the wider natural order. But that is precisely what a theoretical reduction of representational states will do: it clarifies, explores and connects them with the wider natural order; and the development of a theoretical reduction establishes a systematic connection between the representational state and the rest of nature. After all, it provides a reduction of the notion of representation in more basic natural notions which are more widespread in nature; and it also establishes systematic connections between the representation, the natural states that constitute the basis of the reduction and the other natural states which are connected with them in multiple ways.

The final reason is methodological. We should appeal to a supposed naturalist reduction of representation since it would be ontologically more parsimonious to consider the representation as reducible to a natural notion rather than as primitive. After all, the treatment

of representation as a reducible state does not enrich the fundamental states in our ontology.

In sum, even assuming Burge's criterion that the notion of representation should be accepted by a naturalist philosopher because it earns its keep in science by figuring in successful psychological explanations and so that there is no necessity to make it scientifically respectable, there is still a strong motivation to carry on with the reductionist enterprise.[41] This is the positive front of my response to Burge's objection; let's move on to the negative one.

There is a fundamental problem in Burge's defence of the view that there is a motivation for accepting representation as a primitive notion. It is not because successful scientific explanations regard a notion as primitive that it follows that there is a strong motivation for accepting this notion as primitive. That is, the fact that a given successful scientific explanation uses a certain notion as primitive does not imply that it is irreducible. That can even sometimes be the case, but this is not a sound motivation. Notice that what Burge is claiming is that the fact that a successful scientific explanation regards a notion as primitive constitutes a motivation for accepting it as primitive, but not that this is a sufficient reason for a definitive conclusion for its primitiveness. Burge holds that because successful psychological explanations appeal to representation as a primitive notion, there is a motivation to get rid of the reductionist enterprise – the prospects are not good for a viable reduction of representation. In what follows, I will argue *contra* Burge that this fact does not constitute a motivation for primitivism about representational states.

There is a lively debate in biology and philosophy of biology whether classical genetics is reducible to molecular genetics or not. This debate is alive regardless of the fact that classical genetics has successful explanations precisely because this is not a sufficient motivation for treating specific notions of classic genetics as not reducible. The antireductionist's fundamental

[41] Notice that this is a debate on the *motivations* for reductionist naturalism, not on the reasons that purport to demonstrate that representational states are reducible to more primitive natural states. That is, the validity of these motivations entails only that we should engage into the reductionist enterprise. Whether it will succeed in reducing representational states is a further matter that will be decided in the course of this very enterprise.

claim is that any number of different molecular arrangements could correspond to a single notion in classic genetics – *gene*, *locus*, *allele*, etc. Hence, supposed bridge laws for these notions would relate each of these kinds to many molecular kinds and so would not be genuine bridge laws. Notice that this debate is not only about the issue of the incorporation or integration of a reduced theory with a reducing theory, but also about the ontological issue of whether or not the entities posit by classic genetics are reducible to the entities of molecular genetics. If classic genetics is integrated to molecular genetics, then the entities referred to by the specific notion of classic genetics are reducible to the entities of molecular genetics; and so the ontological reduction is fulfilled.[42] Note that this debate is not the only counter-example in biology to Burge's objection, there are also other debates – e.g., whether evolutionary biology is an autonomous discipline or is reducible to molecular biology (ROSENBERG, 2006).

Finally, there is also a counter-example from the history of thermodynamics. There was a lively debate among physicists about the nature of heat between the 18th century and the beginning of the 20th century. Different theories of heat were developed in this period, mainly the caloric theory which explained heat in terms of the flow of a hypothetical weightless fluid called caloric and the kinetic theory according to which heat should be explained in terms of kinetic energy transfer. In the early 1850's, the laws of thermodynamics were established by R. Clausius, W. Thomson and W. Rankine which appeals to the notion of heat. Even after the establishment of the successful laws of thermodynamics, the debate on whether heat is reducible or not to some other notion persisted. In the last decades of the nineteenth century, there was a contentious debate in which on one side E. Mach, P. Duhem and other physicists were objecting to the kinetic theory while L. Boltzmann and others were defending it. Finally, the defenders of the kinetic theory won this debate and it is now established that heat consists

[42] There is an exciting overview of the reductionist debate on classic genetics in "Sex and Death: an introduction to Philosophy of Biology" (STERELNY & GRIFFITHS, 1999).

in the transference of kinetic energy from one body to another via molecular motion. That is, it is now established that heat is reducible to molecular motion, not that it consists in some irreducible substance like caloric. So, the establishment of the successful laws of thermodynamics in which the notion of heat plays a central role happened long before the establishment of heat as reducible to molecular motion. This counter-example shows that in the history of thermodynamics, the establishment of the laws of thermodynamics was not a motivation for treating heat as a primitive notion and instead there was a lively debate on this issue that was only concluded with the ultimate conclusion that heat is kinetic energy transference via molecular motion.[43]

What is the lesson to be drawn from these counter-examples? In order for Burge to defend his objection against reductionist naturalism, he has to show why it is the case that in psychological explanations, contrary to what happens in other branches of science, the fact that a successful scientific explanation appeals to a notion as primitive constitutes a strong motivation for regarding it as primitive. That is, in the case of psychological explanations that regard the notion of representation as primitive, Burge has to show that this constitutes a strong motivation for regarding representational states as irreducible.

Burge could reply, in relation to the above counter-examples from biology, that there is an internal debate among biologists about the viability of the reduction of notions in classical genetics and evolutionary biology. So, they do not constitute genuine counter-examples to his criterion that philosophers should accept that an established scientific notion is primitive provided that this notion appears in a successful scientific explanation as primitive. Therefore, if there is still a lively debate among scientists about its reducibility, then this is not an established scientific notion and so philosophers should remain neutral about this issue, instead

[43] This brief description of the debate on the nature of heat in the history of thermodynamics is based on the first chapter of Stephen Brush's book "The kind of motion we called heat – a history of the kinetic theory of gases in the 19th century" (BRUSH, 1976).

waiting for a scientific consensus about it.

But how decisive is the scientist's word about the irreducibility of a given scientific notion? Is it the final word? I don't think that this is always the case. Because of a divergence of theoretical interests, philosophers may worry about the irreducibility of a given scientific notion while scientists are neutral about this matter simply because they happen to be not interested in it and have never even thought about it. This may be the case with the notion of representation in psychological explanations, as well as with other notions in different areas of science. The second problem with the above reply is that it implicitly assumes that there is a difference of kind between the philosophical activity and the scientific activity, not a difference of degree as assumed by methodologic naturalists. So, there would be an unbridgeable gap between the nature of science and philosophy. It is not my goal here to take a position in this debate, but the onus of argument is on those how hold that there is a difference of kind between philosophical and scientific activities, not to just assume or even to be neutral about it.

The attack on teleosemantics – the mismatch objection

Burge starts his attack on teleosemantics by assessing causal theories and their project of reducing representational states in terms of causal relations, but quickly rejects it by claiming that causal theories fail to give an account of the problem of misrepresentation. So, he introduces teleosemantics as the most promising naturalist theory to solve this problem via the identification of truth conditions with fitness conditions. However, Burge claims that teleosemantics is ultimately untenable; to show this, he developed the mismatch objection (BURGE, 2010, p. 301-3).

The mismatch objection maintains that there is a fundamental mismatch in the identification of truth conditions and fitness conditions which lies at the heart of teleosemantics. It claims that truth conditions are not identical with fulfilment conditions of

biological function. The performance of a biological function by a given system or trait contributes to its fitness and therefore has fitness value. However, Burge objects, a true representation has no fitness value in itself. While it is guaranteed that the performance of a biological function has fitness value, there is no guarantee that a true representation has fitness value. On the one hand, misrepresentation may not be the failure to fulfil any biological function, and on the other hand, a true representation may fail to fulfil any biological function. In sum, representational success is not identical with fitness success.[44] Therefore, representational states are not reducible to biological functions and the teleological enterprise is doomed.

Challenged by this objection, the teleosemanticist replies that the biological function of a representational state is to detect the presence of a given distal condition and that the detection of this is in itself a contribution to fitness, while a failure of detecting it is in itself a failure to contribute to fitness. However, Burge objects, detection in itself is not a contribution to fitness. No biological function resides *strictly* in the detection of anything. Rather, it is the causal properties of the representational state which initiate or trigger the response to the detected distal condition that actually contributes in itself to fitness. That is to say, the representational state contributes to fitness by triggering the response (usually a behaviour) towards the distal condition, not by detecting it. So, the biological function of the representation is to trigger the organism's response to the distal condition. It is this initiation, not the detection *per se* that contributes to fitness.[45] This thesis is well illustrated by the aforementioned case of the frog: the biological function of the frog's representation of the fly is not to detect it, but rather to

[44] "There is, however, a root mismatch between representational error and failure of biological function. The key deflationist [teleosemantic] idea in explaining error is to associate veridicality and error with success and failure, respectively, in fulfilling biological function. […] Fitness is very clearly a practical value. It is a state that is ultimately grounded in benefit of its effects for survival for reproduction. […] But accuracy is not in itself a practical value. Explanations that appeal to accuracy and inaccuracy— such as those in perceptual psychology— are not explanations of practical value, or of contributions to some practical end." (BURGE, 2010, p. 301)

[45] "*in itself* detection does literally *nothing* to contribute to fitness. It is the causal properties of the detecting state in affecting responses that contribute" (BURGE, 2010, p. 301).

trigger the appropriate response towards it – the frog's behaviour of capturing the fly.

However, this argument is flawed and the thesis that the biological function of the representation is only to trigger the appropriate response towards the detected distal condition is untenable. The first problem is that it is not true that the detection by itself does not contribute to fitness while the triggering of the response by the detecting state does. Rather, no effect in the chain of effects which increase fitness by itself contributes to fitness. Consider the case of the frog again. The snap of the frog's tongue does not contribute by itself to fitness either, what contributes by itself is the whole chain of effects – the detection of the fly, its capture, its digestion, the transport of the resulting nutrients in the bloodstream, etc. Notice that the capture of the fly does not contribute to fitness without the digestion of the fly, the transport of the nutrients in the bloodstream, etc. No single effect by itself contributes to fitness, only the wholly chain of effects contributes by itself. The detection is an effect in this chain that contributes to fitness precisely because it is an indispensable element of this chain, just like the capture of the fly or its digestion. A given effect in the chain does not contribute to fitness by itself but does so in virtue of being part of a chain of effects which as a whole contributes to fitness, provided that the absence of this effect would disrupt the contribution to fitness.

Burge's thesis is that the detection does not contribute to fitness by itself. The truth of a given state is not adaptive *per se*.[46] The detecting state, however, has an adaptive effect since it triggers the appropriate response and hence increases fitness. So, the frog's detecting state contributes to fitness by triggering the catching behaviour, not by detecting the presence of the fly. However, this thesis is untenable. In order to see why, it is necessary to see in detail how the detection of a distal condition is an effect of the representational state.

First of all, to say that the biological function of the system is to detect a given distal condition is a *façon de parler.* What is being said is that the function of the representational

[46] "evolution does not care about veridicality [i.e., truth conditions]" (BURGE, 2010, p. 303).

system is to produce a representational state in reaction to the presence of the distal condition and the production of the representation is precisely the first effect of the reaction of the representational system to the presence of the distal condition. The detection of the distal condition by the representational system is the production of the representational state in response to the presence of the distal condition. The biological function of the representational system is to produce representations in covariance with the distal condition (even if this correlation is not perfectly reliable). Another way of stating this is to say that the biological function of the representational system is to produce true representations (after all, the representation is true if and only if the condition that it represents obtains). Thus, the first effect of the reaction of the representational system to the presence of the distal condition is the detection of the representational state, i.e., the production of the representational state. The second effect is the triggering of the response directed to the distal condition (usually, a behaviour). In the case of the frog, the first effect is the production of the representational state which consists in a neural firing in the brain and the second effect consists in the triggering of the behaviour that normally results in the capture of the fly.

Why take the triggering of the response to the distal condition as an effect of the representational system that contributes to fitness but not the detection of the distal condition that consists in the production of the representational state? This is plainly arbitrary. There is no principled way of maintaining that the contribution of the representational system to fitness resides only in the triggering of the response by the representational state but not also in the production of the representational state that consists in the detection of the distal condition to which the appropriate response is directed. The right criterion to establish which effects of the representational system contribute to fitness is every effect whose absence would prevent the response to increase fitness. That is, the neural firing which constitutes the detection of the distal condition, the triggering of the response, etc. Notice that in case of failure in detection or

the absence of the detection itself, there would not be any adaptive response at all. The production of the representation which consists in the detection of the distal condition is a necessary element in the chain of effects, without it there would be no triggering of the response and hence no adaptive response. In the case of failure, there would be no adaptive behaviour at all (the frog would not capture the fly), and in the case of the absence of the neural firing which constitutes the detection of the fly, there would be no trigger of any behaviour. The detection of the fly is an indispensable element in the chain of effects, without it there would be no increase of fitness. Burge could reply that the production of the representation is adaptive only inasmuch as it triggers the response, not because it detects any condition. But notice that in the absence of the detected distal condition, the response would not be adaptive – it would trigger a non-adaptive response.

Another argument in favour of such strong dependence between the representational system's effect of detecting the distal condition and the effect of triggering the response is that subpersonal representations like the frog's are pushmi-pullyu representations.[47] In these cases, there is not an imperative content distinguished from an indicative component as it happens in personal representations. Rather, the representational state simultaneously indicates the presence of the distal condition and dictates an appropriate response towards it. The unique component of the representation is constituted by an indication of the presence of the fly in a certain direction and the dictation to the frog to snap its tongue in this direction. Evidently, one component may be abstracted from another in such a manner that from a theoretical point of view it is possible to distinguish an indicative and an imperative component, but in reality there is no such distinction. So, the connection between the imperative and the indicative component in pushmi-pullyu representations is stronger than in non-pushmi-pullyu representations. The fact that representations like the frog's are pushmi-pullyu representations makes even more

[47] This term was originally coined by Ruth Millikan, cf. MILLIKAN, 1995.

untenable the claim that the representational system's effect of trigging the response directed to the distal condition contributes by itself to fitness but not the detection of this condition.

Burge recognizes that there is a strong coincidence or correlation between the detection of the distal condition (i.e., the production of the representational state) and the triggering of the response directed to the distal condition. However, he notes, even this strong coincidence or correlation is not identity and indeed there are cases in which they come apart. In order to demonstrate this, Burge appeals to real examples in which the representational system fails to detect the distal condition but nevertheless triggers an adaptive response. Thus, there is no sense in saying that in these cases the representational system fails to perform its biological function – after all, there is contribution to fitness. So, Burge concludes, representational systems were selected not because of their accuracy in detecting distal conditions, but because they trigger the responses to the distal conditions.

Burge states that there are a plenty of cases in which representational systems contribute to fitness by trigging a behaviour which increases strength and agility and so ultimately is adaptive even when it misrepresents the presence of predators in the environment. In his own words,

> "suppose that the avoidance mechanism functioned to increase strength and agility — in avoiding the predator — even in cases in which the animal engaged in avoidance behavior, because of a misrepresentation as of a predator, when no predator was present. Suppose that in each case, whether or not the predator is present, the avoidance mechanism contributes to the animal's fitness for avoiding predators. Then, although the ultimate *raison d'être* for the mechanism might be absent in a given case, there would be *no* biological sense in which the mechanism failed to fulfill a biological function when it effected avoidance behavior in cases where the distal condition was not present." (BURGE, 2010, p. 302)

That is, the representational system contributes to fitness no matter whether it accurately represents or not the presence of the predators because it increases strength and agility. So, in

this case the representational system performs its biological function and there is no biological malfunction. This argument, however, is flawed.

The reason is that a given system may have more than one biological function. Evolutionary selection is plainly compatible with distinct and parallel functions. So, in Burge's example, the representational system may have two distinct parallel biological functions - to detect predators and to increase strength and agility. Assuming that the representational system has two parallel biological functions, it is perfectly possible for it to perform its function of increasing strength and agility but not its function of detecting predators and vice versa. It is plainly compatible that the system fulfils one function, while failing to fulfil some other function. So, the representational system contributes to fitness in light of the adaptive effect of increasing agility and strength but fails to contribute to fitness in light of the adaptive effect of avoiding predators. Generalizing this result, in Burge's case the representational system has the biological function of detecting a given distal condition that results in the avoidance of predators and also the biological function of triggering a behaviour which results in the increase of strength and agility.

My thesis is that the detection of the distal condition and the triggering of the response to the distal condition are both adaptive effects of the representational system. On one hand, in the majority of cases they will constitute only one biological function, with the detection of the distal condition causing the triggering of the response to the distal condition, with the failure of detecting the distal condition resulting in no increase of fitness (e.g., in the case of the frog). Notice that both effects are indispensable and that the absence of one of them will result in a break in the chain of effects which ultimately result in the increase of fitness. On the other hand, in minority cases like Burge's one, the representational system has two parallel biological functions respectively constituted by the adaptive effect of detecting the distal condition to avoid predators and the adaptive effect of triggering avoidance behaviour to increase agility

and strength.[48] In these cases, the representational system has the function of detecting the distal condition to avoid predators no matter whether this behaviour will increase strength or not, and it also has the function of triggering the avoidance behaviour to increase strength and agility no matter whether this behaviour will avoid predators or not. Both functions are parallel and compatible and, as a matter of fact, nature is rife with biological systems that were selected to have more than one adaptive effect and thus more than one biological function. Finally, notice that only the first function has a semantic nature, the second function has no semantic nature at all – it has nothing to do with the representation of anything. It is constituted by the effect of generating a certain behaviour that increases agility and fitness, but the triggering of this behaviour is independent of the representation of any distal condition. In fact, there are several examples of systems that also increase agility and strength but which do not represent at all. It is just a coincidence in these minority cases that the system that contributed to the increase of strength and agility by triggering this kind of behaviour is also a representational system - it might not have been.[49]

It will be useful to compare the above response to Burge with responses to another problem facing teleosemantics. The problem is that, since there are cases of representations which really serve a biological function in virtue of being false, truth conditions cannot be identical with fitness conditions and thus teleosemantics is flawed. For instance, there are cases of depressive realism in which psychologically healthy people tend to have inflated beliefs about their own social status, in opposition to depressed people who tend to have accurate beliefs about their own social status. Let's assume that these false beliefs among healthy people

[48] Or even a third biological function constituted by a third effect as long as it is adaptive. There is no pre-established limit of the number of parallel functions that a given biological system may have.

[49] It could be replied that this argument developed here does not really address Burge's objection because he is actually objecting that the detecting state is not genuinely a representation, but just a sensory state with no truth-condition. In fact, he objects that teleosemantics is too liberal since it treats certain states that clearly are not representational as representations. But this is a distinct objection and I will not address the objection of liberality here, but only in the next chapters since it is part of a bigger issue, the problem of demarcation.

have the biological function of encouraging them to be enterprising by increasing self-esteem. Thus, contrary to teleosemantics, these beliefs are adaptive in virtue of being false, not true.

The standard reply to this problem consists in stating that this objection only arises because it mistakenly assumes that these false beliefs have only one biological function, when in reality it has more than one function (PAPINEAU, 2016). They have the function of accurately representing reality and also the function of encouraging being enterprising. So, they fail to perform the function of accurately representing since they are false but they fulfil the function of encouraging enterprise.

In effect, I adapted this standard reply in explaining why we should not worry about Burge's case. However, by doing this I am not committed to this standard response to the objection about depressive realism. These are different and independent cases and indeed I think that this two-functions strategy is more promising in order to deal with Burge's case than to deal with depressive realism case. The reason is that in Burge's case it is a matter of coincidence that the same representational system has the function of producing accurate representations and the function of increasing strength and agility, while in the depressive realism case it is not a coincidence that the same representational system has the function of producing true beliefs and the function of producing false beliefs in order to encourage enterprise. The crucial difference between these cases is that in Burge's case it is a coincidence that the production of a false representation increases strength and agility, while in the depressive realism case the belief is adaptive in virtue of being false – only a false belief can have this consequence. After all, only a representational system can produce false beliefs in order to encourage enterprise, but other systems may increase strength and agility by non-representational means.

So, in the depressive realism case there is the problem of explaining how a representational system can be adaptive in virtue of producing sometimes true beliefs,

sometimes false beliefs. Furthermore, how can they constitute distinct functions given that their functional statuses are not completely independent (in depressive realism cases, it is impossible for the representational system to simultaneously fulfil both functions of accurate representing reality and encouraging enterprise)? By contrast, these problems do not arise in Burge's case since here it is just a coincidence that sometimes the production of a false representation will be adaptive and the relevant functions are completely independent – it is plainly possible for them to be simultaneously fulfilled or not fulfilled, or one fulfilled but not the other and vice-versa.

After developing this proposal in order to deal with Burge's case, I have since discovered that Agustin Vicente has also adopted the same strategy: the representational system has the biological function of detecting predators and also the function of increasing strength and agility (VICENTE, 2012). However, Vicente's proposal is fundamentally different from mine. He claims that they are biological functions of different kinds. The function of increasing strength and agility is an *aetiological* biological function, but the function of detecting predators is a *non-aetiological* biological function. The only adaptive effect of the representational system is the effect of increasing agility and strength. Evolution does not care at all about the accuracy of the representational system – it does not care whether or not the system detects the distal condition. Rather, he argues, the representational system has a non-aetiological biological function of detecting the distal condition because it is a "special kind" of biological function. But then why does the representation system have this biological function given that it was not selected for detecting the distal condition? According to Vicente, because it is "the result of a process of natural selection". He concludes that "this may take us to reconsider the selected effect account of functions [i.e., the aetiological conception of biological function]" (VICENTE, 2012, p. 132).

Burge and Vicente agree that the representational system's effect of detecting the distal

condition is not adaptive and as a result was not selected by evolutionary selection. Thus, it is not an aetiological biological function of the system. However, while Burge claims that accurate detection is not at all a biological function of the representational system, Vicente claims that it is a biological function, though not an aetiological one, which leads him to propose the revision of the aetiological conception of biological function. By contrast, what I have being arguing here is that the biological function of detecting the distal condition and the function of triggering the response to the detected distal condition are *both* aetiological biological functions of the representational system. So, my disagreement with both Burge and Vicente is that they claim that the detection of the distal condition is not an adaptive effect of the representational system, while I am arguing that it is an adaptive effect.

The problem with Vicente's proposal is that it is completely obscure what kind of special biological function the representational system's function of detecting the distal condition is. Notice that it is incompatible not only with the aetiological conception of biological function, it is also incompatible with other available conceptions of biological function such as the systemic conception (CUMMINS, 1975). So, it seems that Vicente would have to develop a completely new conception of biological function to accommodate his thesis that the accurate detection of the distal condition is a distinct kind of biological function. There is no problem with alternative conceptions of biological function and I am open to pluralism about the nature of biological functions. However, since Vicente has not clarified the nature of his new conception of biological function, it is not possible to assess its viability or the thesis that the detection of the distal condition is a special kind of biological function. My conclusion is that until the nature of this new conception of biological function is clarified, the thesis that accurate detection is not an adaptive effect of representational system but nevertheless is a special kind of biological function is implausible and unpromising.

Conclusion

In this chapter, I have first introduced intentional naturalism and its reductionist and primitivist versions. I then presented motivations for reductionist naturalism in contrast with primitivist naturalism. After this, I introduced my favoured teleological approach, i.e., teleosemantics. Finally, in the last section I rejected Burge's primitivist challenge to reductionism; that is, his attack that the motivations in fact support primitivist naturalism, not reductionism. I have also rejected the mismatch objection to teleosemantics. Burge also presented another objection to teleosemantics, the objection of liberality. However, I have not addressed this so far, as behind the objection of liberality lies a much larger issue: what are the minimal conditions for intentionality required to distinguish genuine representational states from non-representational ones? This is the problem of demarcation. In the next chapter, I address this problem and reject some responses to it, including Burge's own demarcation proposal. In the fourth chapter, I present and defend my own solution to the problem of demarcation – the dual proposal.

CHAPTER 3. THE MINIMAL CONDITIONS FOR INTENTIONALITY: THE PROBLEM OF DEMARCATION

What are the minimal conditions for intentionality that a given state should satisfy for it to constitute a representational state? That is, what are the limits of intentionality? This is the problem of demarcation. My approach to this problem will be divided into two chapters. In the first section of the present chapter, I introduce the problem of demarcation and show its relevance for the debate on the viability of naturalist theories of mental representation. In the second and third sections, I assess methodological and substantiality worries related to this problem. I first reject the view that the problem of demarcation is just terminological; after that I propose the adoption of a variation of the method of reflective equilibrium to develop and assess proposals for minimal conditions for intentionality. Finally, in the two last sections of this chapter, I assess and reject two demarcation proposals for the limits of intentionality: the causal independence and constancy mechanism proposals. The next chapter is dedicated to the positive side of my approach in which I will develop my own proposal for minimal conditions for intentionality – the dual proposal.

3.1 The problem of demarcation

Naturalist theories of mental representation are often criticized for being either too liberal or too restrictive about the requirements for a given state to be a representational state. A theory of representation is *too liberal* if it treats certain states as representational states when they are clearly not representations; it is *too restrictive* if it treats certain states as non-representational states when they are clearly representations. Behind both objections, lies the problem of demarcation: what are the limits of intentionality? What is the border of intentionality that distinguishes the limiting cases of representational states from non-representational states? The objection that a theory is too liberal is just the objection that it has drawn the limit of representationality too low, while the objection that the theory is too restrictive is just the objection that it has drawn this limit too high.

In the case of teleological and causal theories of content, the 'liberal' side of the problem of demarcation raises an objection. They are often challenged for treating states that clearly are not representations as representational states – they draw the limit of representationality too low.[50] The following cases are illustrative of this objection: magnetosome states of anaerobic bacteria and the reptile's body states which varies with the heat of the sun (BURGE, 2010, 300; 303-4); hormone concentrations in our blood; detection of light in amoeba or planaria for phototaxis (SCHULTE, 2015, pp. 119-20; FODOR, 1986, p. 10-11); etc. For a variety of reasons, it has been claimed that it is a mistake to treat these states as genuine representations, i.e., that they clearly do not represent anything. Thus, the conclusion that these naturalist theories should be ruled out or at least highly refined.

But how to determine which states are genuine representations and which ones are not? A useful way of approaching this issue is to ask what are the *minimal conditions* that a given state should satisfy for it to constitute a representational state. Minimal conditions for

[50] This is also called "the problem of the breadth-of-application" (BURGE, 2010, p. 304; SCHULTE, 2015).

intentionality are those conditions satisfied by states in the lower level in the hierarchy of representational states. If a given state does not satisfy one of these conditions, it follows that it lies outside the representational realm – no matter how it might look like a representation.

The hierarchy of representational states is well illustrated by checking the states which constitute it. There are very simple and primitive representational states at the bottom of the hierarchy, very sophisticated representations at the top and representational states with intermediate sophistication at intermediate levels. That is, the hierarchy is constituted by higher-level representational states at the top, lower-level representations in the bottom and intermediate-level representations in between. At the top there are representations like desires, beliefs and other propositional attitudes; abstract and scientific representations (e.g., representations of numbers, quarks etc.); and so on. At intermediate levels, there are less complex representations like pre-linguistic infant intentional gestural signals (BATES et al., 1975), great ape gestural communications (CALL & TOMASELLO, 2007), and so on. What about the primitive states at the bottom of the hierarchy? We can ask ourselves a number of subsidiary questions about primitive states: what kind of states are there? Are honeybee dances which indicate the whereabouts of nectar genuine representations? Are representational anaerobic bacteria's magnetesomes states which indicate the direction of the magnetic field and hence the direction of the bottom of the ocean? Are amoebas and planarias representational systems? To answer these questions, one must establish minimal conditions for intentionality, based on which one can determine which representational states lie at the bottom level of this hierarchy. That is the only way to have a justified response for the problem of demarcation. Otherwise, any response would be arbitrary and unfounded.

Consider a group of representational states. Some states are higher-level representations whose representational statuses are uncontroversial while other states are primitive states whose representational statuses are controversial. Now suppose that there is a demarcation line

distinguishing representational states from non-representational ones. Accordingly, the closer a given representational state is to this line, the more doubtful is its representational status. This line also distinguishes primitive representational states from other states that even though are not representational, their non-representational statuses are also controversial. These are borderline cases of primitive representational states. The problem that arises is how to demarcate which states are representational and which states are not. It is here that minimal conditions for intentionality come into play. Their role is to establish conditions to distinguish, among borderline cases of primitive representations, states that are genuinely representational from states that are not.

Here I assume that primitive representations are sensory representations such that the limits of intentionality are demarcated by sensory representations. That is, lower-level representational states *are* sensory representations. The role of minimal conditions for intentionality is to distinguish sensory representational states from non-representational states. The latter just register or indicate the relevant conditions, in contrast with sensory representations that represent them. If paramecia are representational systems, they represent the direction of light; if they are not, they merely indicate or register it. Non-representational sensory states have sensory discriminations of the relevant conditions, but there is no genuine representation, in contrast with sensory representations that genuinely represent them.[51]

There are several proposals for minimal conditions for intentionality. It is not possible to assess all of them, so I will limit my assessment to two of them. The reason for this choice is that, as it will become clear later, both proposals contribute to the development of the proposal that I develop in the next chapter. First, I assess the causal independence proposal.

[51] The assumption that primitive representational states are sensory representations is widespread in this debate. Millikan claims that pushmi-pullyu representations are primitive and that they are sensory representations (MILLIKAN, 2004, p. 158). Burge and others claim that the lower border of perception is the lower border of intentionality (BURGE 2010, p. 317; SCHULTE, 2015). Here I will not commit myself to either of these claims, but only with the weaker claim that primitive representational states are sensory representations.

Jerry Fodor and Ansgar Beckermann have proposed that a given system does not count as genuinely representational unless there is a *causal independency* between the tokening of the relevant external condition and the tokening of the state which supposedly represents it. The minimal condition for intentionality is that there is a causal independence between the presence of the relevant stimulus and the tokening of the state (BECKERMANN, 1988; FODOR, 1986). Second, I assess *the constancy mechanism proposal*. Tyler Burge and Kim Sterelny have proposed that it is a minimal condition for intentionality of a given state that the system employs a constancy mechanism on its production (BURGE, 2010, pp. 315-9; 342-436; STERELNY, 1995). Before assessing these proposals, in the next two sections I give an account of some methodological and substantiality worries related to the problem of demarcation.

3.2 A terminological problem?

It is a common concern whether what is at stake on the problem of demarcation is just a terminological issue about what is called "representation". Here is the line of reasoning: it does not matter whether you call a given state a "representation" or not, for it does not make any difference. For instance, you may call the amoeba's state that covaries with light to enable phototaxis whatever you want – a "representation", a "registration" or an "indication" of light. What difference does it make? Whether it is appropriate or not to call it "representation" is just a terminological issue. What we call this state does not make any explanatory difference for the success of scientific explanations of the nature of amoebas. The same goes on in the case of controversial representational states of bacteria, honeybees, and so on. The conclusion is that the problem of demarcation is a *scheinproblem*, a pseudo-problem devoid of any real substance.

How should this concern be assessed? At first glance, this objection may look plausible,

but it misses the real nature of the problem of demarcation. Evidently, you can call anything by whatever name you want – "representation", "registration", "indication", etc. Indeed, it makes no difference which terminology is adopted, but what you mean by the term you employ makes a difference. The meaning assigned to your chosen terminology is essential: if by "representation" you mean what is commonly meant by "representation" in English, then the question of whether a given state is a representation or not is substantial. That is the case because the notion of representation is pre-theoretic: it precedes any scientific or philosophical investigation on the nature of representational states. This fact by itself establishes constraints on which states genuinely represent other states and which do not. Since the notion of representation is pre-theoretic, the question of whether a given state is a genuine representation of another state gives rise to intuitions into our minds that may go in the direction of accepting or rejecting it as a genuine representation. These pre-theoretic intuitions constitute a constraint on what should be properly considered a genuine representation to a certain degree. That is one reason why several theories of representation are accused of being too liberal or too restrictive depending on what states they consider representations. If a given theory takes certain states to be genuine representations and our intuitions indicate that they are clearly not representations, then this theory will be sooner or later accused of too liberal; while if this theory does not consider certain states as representational and our intuitions indicates that they are clearly representational, then this theory will probably be accused of too restrictive. That is how this game is played among philosophers of mental representation. Not surprisingly, that is when the problem of demarcation arises among them.

But first, why should our pre-theoretic institutions constitute a constraint on what successful scientific theories of cognition should appropriately posit as representational states? The most popular view among philosophers of mental representation nowadays is precisely that the distinctive explanatory role played by the posit of representational states in successful

scientific theories is a reason good enough to justify even the representational status of states that we intuitively don't take to be genuinely representational. What better reason could we have to accept a counter-intuitive representational state than that it plays the distinctive explanatory role in a successful scientific theory of cognition? After all, the notion of representation is a technical and scientific notion. In short, the requirement that the distinctive explanatory power of positing representational states is the supreme constraint for assessing whether a given state is genuinely representational or not.[52] Furthermore, one of the lessons that philosophers usually take from the enterprise of naturalizing intentionality is that intuitions should be put aside because (among other reasons) intuitions commonly go in the opposite direction of the results of successful scientific theories. This explanatory requirement view constitutes a serious challenge to the thesis that pre-theoretic intuitions constitute a constraint on assignments of representational statuses.[53]

It is true that our pre-theoretic intuitions may clash with successful scientific theories of cognition on the matter whether the posited representational states are genuinely representational. For instance, is it intuitive to claim that rats genuinely represent the local environment?[54] When doing their jobs, philosophers of mental representation should certainly not put too much weight on what pre-theoretic intuitions dictates about mental representations. That is, our pre-theoretic intuitions cannot have the *final* word when engaged in on a philosophical investigation on intentionality. However, they cannot be entirely thrown away either. That is the case because if a given scientific explanation posits a state as

[52] This explanatory view is highly connected with the explanatory requirement that constitutes the basic conception of mental representation in the sense that the explanatory requirement is used there to specify what is a mental representation, while on this view it is used to establish the limits of intentionality.

[53] "The project is to characterise such putative intentional properties by examining the theoretical role they play. From this perspective, pre-theoretic intuitions about cases have no special status." (SHEA, 2007, p. 406).

[54] Place cells on rat's hippocampus fire when and only when the rat is in a certain location in the environment. John O'Keefe and Lynn Nadel have proposed that the hippocampal formation functions as a cognitive map consisting of a set of place representations (Cf. O'KEEFE & NADEL, 1978). This is the "hippocampal representation" of local environment (O'KEEFE & BURGESS, 2005, p. 855). But one could claim that it is not intuitive that rats are intentional systems because they lack the required cognitive complexity to be intentional.

representational, but it has nothing to do with what we ordinarily understand as a representation, then this state is not a representation at all. Evidently, the scientist may decide to baptize the posited state in the scientific explanation with whatever preferred term, but it does not touch in any sense the conclusion that this state is not a genuine representation, since it is entirely incompatible with our intuitions on the nature of representational states.

Finally, the last reason for the substantiality of the problem of demarcation arises from its comparison with other philosophical problems. Note how many traditional and substantial philosophical problems would be dismissed on the grounds that they are easily solved by just providing new definitions of the relevant terms and putting aside our pre-theoretic intuitions on this matter. For instance, consider the question: can computers have consciousness? All that you have to do is to redefine "consciousness" as "electrical activity", put aside your intuitions contrary to the claim that mere electrical activity is enough for consciousness and then the problem is solved. Evidently computers have consciousness, for they are machines with electrical activity. Consider the trolley problem: what is the morally right thing to do? Should we let the trolley kill five people on the main track or should we pull the lever, putting the trolley in the sidetrack and killing just one person? If one redefines "moral" as a feature of every human action and puts aside the moral intuitions, then both actions are equally morally right. The dilemma is dissolved. Evidently, such a strategy is a misleading way of answering these questions, since it does not address the genuine philosophical problems behind them – respectively, the nature of consciousness and the moral dilemma behind the thought experiment. Likewise, the adoption of this strategy to deal with the problem of demarcation does not *really* address the philosophical problem behind it – the nature of intentionality in general and the minimal conditions for intentionality in particular. This is not a terminological problem precisely because our intuitions have a role to play on the assessment of posits of representational states by cognitive theories. The notion of representation is not just a technical

notion that has nothing to do with our intuitions and so the its comparison with purely technical notions like "quark" or "gene" is fully misleading.[55]

Notice that I have defended the view that intuitions constitute a constraint on what are properly posited as representational states by cognitive theories. Intuitions, however, do not constitute the final word on this matter; and yet they cannot be thrown away, either. However, I have not committed myself to the extent in which pre-theoretic intuitions constitute a constraint. Such commitment is not necessary to show that the problem of demarcation is substantial. The problem of demarcation is substantial, not terminological. The urgency of investigating the minimal conditions for intentionality still stands. But before engaging in this investigation, it is necessary to carry out a previous investigation on the status of our pre-theoretic intuitions on the assessment and development of proposals for minimal conditions for intentionality, as well as their relationship with the explanatory requirement.

3.3 The method of reflective equilibrium and the status of pre-theoretic intuitions

I have defended the thesis that the problem of demarcation is a substantial problem about intentionality by arguing that our pre-theoretic intuitions constitute a constraint on the assessment of the positing of representational states by cognitive theories. But what is the status of our intuitions on the development of a solution to the problem of demarcation? Which role should they play in the assessment of minimal conditions for intentionality? And finally, how they are related to the explanatory requirement for the positing of representational states? That is the methodological problem that will be addressed in this section. In what follows, I propose the adoption of a variation of *the method of reflective equilibrium* to develop a proposal for the limits of intentionality, constituted by, on the one hand, mutual adjustments between pre-

[55] William Ramsey has presented similar arguments as the ones above to show that pre-theoretic intuitions are a constraint for the assessment of proposals of the limits of intentionality, cf. RAMSEY, 2007, pp. 11-3.

theoretical intuitions on representationality and, on the other hand, the explanatory role of positing representational states.

First of all, let's address another challenge against appealing to pre-theoretical intuitions on philosophical debates about representations. The challenge runs like this: our intuitions should play no role in the demarcation of the limits of intentionality because ordinary people do not have a pre-theoretical notion of representation; if so, there is no intuitive view about representation in general. People have pre-theoretical intuitions about specific cases – e.g., it is intuitive that beliefs and meaningful words are representations and it is counter-intuitive that stones and clouds are representations. But people do not have intuitions about representation in general. The notion of representation is not a common-sense notion shared by ordinary people. So, it makes no sense to say that the positing of a state as a representation violates our pre-theoretical notion of representation simply because there is no such thing as an intuitive notion of representation.

I doubt that this is the case; rather, it strikes me that there is a pre-theoretical intuitive notion of representation. People have, for example, an intuitive notion of representation that embraces representations of very different kinds: written words, mental states, paintings, etc. But to refute the above argument is not necessary to go as far as asserting that people have a pre-theoretical notion of representation *qua representation*. It is enough to assume that people have intuitions about specific cases of representations – about a written word, a mental state, a painting, etc. The collection of all intuitions about these specific cases of representations shared by ordinary people constitute an intuitive notion of representation even if ordinary people are not aware of their intuitions about representation *per se*. This notion is precisely constructed out of all features that are shared by all instances of representation that people face in their everyday lives. Hence, there is a pre-theoretical intuitive notion of representation in general which is elicited by the intuitions that people have about specific representations.

Having dispelled this worry, I address the two fundamental methodological problems on the status of pre-theoretical intuitions on the problem of demarcation. First, to what extent should our pre-theoretic intuitions about the limits of intentionality be vindicated in the assessment of proposals for the limits of intentionality? In the previous section, I maintained that our intuitions should not be thrown out of the assessment of representational status of posited states by cognitive theories, nor constitute the final word about it. The first problem is how to determine the intermediate level between these two extremes on which intuitions constitute a constraint to assess proposals for the limits of intentionality.

Now let's move to the second methodological problem about the status of our intuitions on the limits of intentionality. Suppose that after this investigation, we have reached a definitive response to this first methodological problem which establishes that intuitions constitute a constraint in the assessment of the positing of representational states to a certain degree. But what if our pre-theoretic intuitions about the limits of intentionality are incompatible? That is, what if the intuitive conception of representation is inconsistent? It is possible that we have incompatible intuitions on whether a given kind of state is genuinely representation or not. If that is the case, then no matter how plausible is the degree in which our intuitions should constitute a constraint to the assessment of demarcations of the limits of intentionality, this constraint must be put aside in virtue of its internal inconsistency. After all, it is a basic principle that appropriate constraints for the assessment of the positing of representational states should be consistent.

The goal of this section is to give a solution for both methodological problems. First, I argue that there is no general rule to determine in advance the degree in which our pre-theoretic intuitions constitute a constraint on the assessment of demarcations of the limit of intentionality; rather, I will propose a variation of the method of reflective equilibrium applied to the case of minimal intentionality. Second, I argue that it is possible that intuitions on the

limits of intentionality are inconsistent, but this is not a reason to conclude that they cannot constitute a constraint. In conclusion, to appeal to Wilfrid Sellars' distinction, the application of the method of reflective equilibrium will result in the establishment of a demarcation of the limits of intentionality which is constrained by the manifest image of representation – our pre-theoretic intuitions – and the scientific image of representation – the successful cognitive theories which posit representational states to explain behaviour (SELLARS, 1962).

In light of the implausibility of defending one of the extreme positions on the degree to which our intuitions constitute a constraint on a demarcation of the limits of intentionality, a natural proposal would be to defend the intermediate position that this demarcation should vindicates only the pre-theoretic essential intuitions. That is, for a given state to have minimal intentionality, we should respect some essential or indispensable intuitions on what is a representation. An essential pre-theoretic intuition is precisely the one which a state should respect for it to constitute a genuine representation. However, this is not a viable proposal.

Suppose that you have a set of all pre-theoretic intuitions on minimal intentionality and that you are wondering whether a given state is representational. How should we distinguish essential from the nonessential intuitions? This is a fundamental problem for this proposal. It is hard to determine which intuitions are essential in order to distinguish them from non-essential ones. Evidently, the determinacy of essential intuitions should be principled, based on a certain criterion. Otherwise the selection of essential intuitions would be plainly arbitrary. But it is hard to conceive of a non-arbitrary criterion that satisfies this condition.[56]

It is clear that some intuitions are more entrenched than others in our intuitive conception of representations. The more an intuition is entrenched, the more weight it should

[56] Furthermore, it is possible that two competing theories posit two distinct kinds of representational states in such way that both states respect all but one of the supposed essential intuitions. It would be absurd to claim that notwithstanding the explanatory power of positing these representations and the fact that they respect all but one intuition, they should not be considered as representations because they violate an essential intuition. It is utterly implausible to claim that a supposed essential intuition has this definitive and conclusive power.

have on the assessment of proposals for the limits of intentionality. But even among the intuitions that are the most entrenched ones, none of them should be necessarily vindicated by a theory of cognition for the posited representational states to be representational. There is always the possibility that we should consider the posited states as genuinely representational even though they do not respect one of our most entrenched intuitions about representations. For instance, we should consider them as representations because they respect other intuitions which are equally entrenched (as well as other less entrenched intuitions) and because they play the relevant explanatory role. All things considered, the explanatory role of the posited representation may compensate for its violation of a highly entrenched intuition. But how can intuitive and explanatory considerations be put together to develop a demarcation proposal?

I think that the only way is via mutual adjustments between intuitions on minimal intentionality and the distinctive explanatory role of positing representational states by successful cognitive theories. The constraints constituted by intuitions and the explanatory role should be both used to assess proposals for minimal conditions for intentionality. On one side, the positing of a representational state can adjust the intuitive conception of representation such that we can start accepting certain states as genuine representations that we were not disposed to accept before. On the other side, the intuitive conception of representation can adjust what states should be properly posited as genuine representations such that certain posited states are no longer accepted as genuine representations. What I am proposing is a variation of the method of reflective equilibrium that was made famous by Nelson Goodman and John Rawls when they applied it respectively to the justification of laws of inference in logic and of principles of justice.[57] My strategy is to adopt a variation of the method of reflective equilibrium to develop the appropriate proposal for the limits of intentionality and to assess rival proposals.

It is highly unlikely that there is any general rule or principle based on which one can

[57] Cf. GOODMAN, 1979, pp. 62-66; RAWLS, 1999, pp. 17-9; 40-46. For a general overview, cf. DANIELS, 2016.

establish the right degree of the extent of which our intuitions on minimal intentionality should be respected in the assessment of responses to the problem of demarcation. This is a doomed enterprise with no prospect of success. It is more viable to assess each proposal to find out which intuitions are vindicated by them, which ones are not and finally to balance it with the explanatory role played by the representational states in the theories that posit them. But how to balance them? The method of reflective equilibrium is the best way of accomplishing this task.[58]

Reflective equilibrium is the method of going back and forth between our judgments about particular cases of a subjective matter: the principles that we accept as governing these judgements and the theories that we accept as explaining these particular cases and which were also based on these judgments and principles. So, in this method, there are three distinct but connected elements playing a role, i.e., *theories*, *intuitive principles* and *judgements*. The three elements should be revised to achieve coherence among them. There will be reflective equilibrium when the revision reaches a state of acceptable coherence. Judgments, principles and theories are justified by being brought into agreement with each other via mutual adjustments. For instance, intuitive principles are amended if they yield judgments that we are not willing to accept; while judgments are rejected if they violate principles that we are not willing to amend. The goal of the method of reflective equilibrium is to achieve a state of optimal equilibrium, that is, a state in which we are no longer inclined to revise none of the judgments, principles and theories since together they have the highest degree of acceptability.

So far, so good. But how should the method of reflective equilibrium be applied to the problem of demarcation? First, we need to specify the intuitive judgments and principles that

[58] Ramsey does not propose the adoption of the method of reflective equilibrium which consists in mutual adjustments between intuitions and theoretical virtues of positing representational states. But he also claims that theories that posit representational states should not be tested by our intuitions based on a general principle that determines in advance the extent of which intuitions constitute a constraint to these theories. Cf. RAMSEY, 2007, p. 8-14.

play a role in this method when applied to the problem of demarcation. After that, we need to specify the theories that posit representational states to explain behaviour.

The relevant intuitive judgments are judgments about whether specific states are intuitively representational or not. We share the intuition that certain states are representational (e.g., beliefs, desires, sentences, perceptions, etc.) and that other states are not representational (e.g., stones, clouds, energy transference, etc.). These are examples of intuitive judgments about the representationality status of specific states. The intuitions are stronger in the cases of conscious representational states like beliefs and desires, while in the case of subpersonal states (those that pertain not to the person but to a part of the person) like the (supposed) representation of edges in V1 of the cerebral cortex posited by neuroscientists are less intuitively representational.

Intuitive judgments give rise to intuitive principles that a given state should respect to constitute a representational state. The recognition that several states that are intuitively representational share a given property gives rise to an intuitive principle. The requirement of the satisfaction of this property constitutes a principle that a candidate for representational state should respect. However, the justification of intuitive principles and judgments goes in both ways because there is mutual justification – certain principles justify certain judgments because they respect these principles; certain judgements justify certain principles because they give rise to these principles. In what follows, I propose the autonomy, complexity and activity principles. I then show how certain judgments give rise to them and how they justify certain judgments. However, this is not an exhaustive list. Rather, these are the principles that I take to be the most entrenched in our intuitive conception of representation. In the next sections I assess the intuitiveness of proposals for minimal conditions for intentionality in light of these three principles.

But before presenting the intuitive principles on the limits of intentionality, it is worth

clarifying one issue. In the first chapter, I have proposed two intuitive requirements that constitute a basic conception of mental representation – the misrepresentation and original intentionality requirements. Evidently, they also constitute the intuitive conception of representation. However, here I am considering only those intuitive principles that concern the limits of intentionality. They are necessary in order to develop and assess proposals for minimal conditions for intentionality. That is, conditions that distinguish, among borderline cases of primitive representations, states that are genuinely representational from non-representational states. By contrast, in the first chapter I have appealed to the misrepresentation and original intentionality requirements to construct a basic notion of what is a mental representation. The problem of the limits of intentionality presupposes a basic conception of representation based on which one can ask what is the lower border of the states that satisfy this basic notion. In sum, the intuitive requirements proposed in the first chapter are concerned with the nature of mental representations, while the intuitive principles that I will propose here are different because they are concerned with the limits of intentionality. That said, let's present these intuitive principles.

First, there is the *autonomy principle*. A genuine intentional system is an autonomous system – the forces responsible for its behaviour originates within the system, not outside. That is, intentional systems are self-moving systems. But why is this principle intuitive? This principle is present in our judgment that a needle which moves in the direction of the magnet is not representing the presence of the magnet in that direction. The needle is not an autonomous system, the magnetic forces which causes its movement in the direction of the magnet comes not from within the needle, but from outside, i.e., from the magnet.[59] By contrast, humans, dogs, bees and other organisms are autonomous systems. That is one of the reasons

[59] To say that the forces responsible for the behavioural output originate within the system precludes an external force to have any influence on its behaviour? Not at all. What matters is that in case of the absence of internal forces, there would be no behaviour at all.

that their intentional status is intuitive. If the behaviour of the organism is generated from forces outside it, then this is not an intentional system. It is a condition for intentionality that the way how the organism behaves is not caused from forces that originate outside the system (BECKERMANN, 1988, p. 127). The autonomy principle partly justifies the intuitive judgements that certain systems are intentional. If it is not intuitively clear that a given sensory system is intentional or not (e.g., planarians or paramecia), the recognition of its autonomy tips the balance in favour of the judgment that the relevant state is representational.

The second intuitive principle is the *complexity principle*. Systems without a certain cognitive complexity are not genuinely intentional. This intuition is highly entrenched in our intuitive notion of representation. The intuitiveness of this principle is grounded on the fact that it is implausible to claim that unicellular organisms like paramecia and other systems are intentional because they lack the minimal complexity required for intentionality. By contrast, it is plausible to claim that humans, dogs and other animals are intentional systems precisely because they have this minimal cognitive complexity. But what is the exact nature of this minimal cognitive complexity? This is an open question. There is no clear-cut line based on which it is possible to establish the minimal cognitive complexity that a certain system should have for it to be eligible as intentional. Nevertheless, the fact that this is an open question does not entail that it is not intuitive that systems without a certain complexity are not representational. It does not follow from the fact that there is no strict limit for the extent of this cognitive complexity that there is no limit at all. Hence, the complexity principle still stands. Finally, just like what happens with the autonomy principle, the complexity principle (partly) justifies the intuitive judgements that certain systems are intentional. If it is not intuitively clear that a given system is intentional, the recognition that this system has a certain cognitive complexity makes it more intuitively representational.

Finally, the third principle is the *activity principle* (BECKERMANN, 1988, p. 130).

Consider a system that has a certain cognitive complexity and the forces responsible for its behaviour are internal. Nevertheless, the system is deprived of any active role – it is wholly passive. The relevant state is just automatically triggered by the presence of the external feature and the state automatically triggers the behavioural output. But what is the distinction between active and passive systems? Paradigmatic cases of passive systems are tropistic systems like amoebas and paramecia (i.e., systems that moves in a certain direction in a direct response to a given external stimulus). The passivity comes from the fact that their behavioural output is *automatically activated* by the respective specific stimuli. But why is the activity principle intuitive?

I think that this principle is intuitive because the notion of one state representing another state carries with it the idea that representing is not just to receive input stimuli from some external feature or just to be in some basic relation with it (isomorphism, correlation, etc.). That is why it is not intuitive that the mercury volume represents temperature even though it receives input stimuli from the environment and there is a strong correlation between the mercury volume and the temperature when the temperature increases or decreases. In the same vein, it is counter-intuitive that a random drawing in a sheet of paper represents a galaxy far away (or *vice versa*) simply because there is an isomorphic relation between them. Behind these judgments lies the intuition that representing is an activity of an intentional system. The mere covariation or isomorphism requires no active role from the system. For one state to represent another is for it to play an active role. That role may be fulfilled in different ways, but some active role is always required. Finally, the activity principle (partly) justifies the intuitive judgements that certain systems are intentional. If it is not intuitively clear that a given system is intentional, the recognition that this system plays an active role makes it more intuitively representational.

But what is the difference between the activity and autonomy principles? Are they

really different? It is possible for a given system to be autonomous but play no active role, but an active system cannot be not autonomous. That is, every active system is autonomous, but there are autonomous systems that are not active. If the system plays some active role, then its behavioural output cannot be entirely caused by forces outside the system. After all, if there is no relevant internal force, no active role may be played. It is required for the system to have some active role that its internal forces are (partly) responsible for the behavioural output. For the system to make any difference for the resulting behaviour, it should have some force to affect the behaviour and that is not possible in case of the behaviour is fully generated by forces outside the system. In sum, the activity principle is more demanding than the autonomy principle. Let me illustrate it with a system that is autonomous but wholly passive.

Paramecia are unicellular ciliated organisms that are highly sensitive to the presence of light: whenever there is light in the surrounding environment, they move in the opposite direction of it. This phenomenon is called tropism, i.e., the turning of the organism (or parts of it) in a given direction in response to the presence of a certain external stimulus. Paramecia and other autonomous are self-moving systems: the forces responsible for the paramecium moving away from light are not (wholly) originated outside the system (e.g., in the light). Nevertheless, tropistic systems are wholly passive: they do not have any active role in the resulting avoidance or approaching behavioural output towards the relevant stimulus. That is the case because tropistic systems move as long as the relevant stimulus is present, but when it ceases, the system immediately stops. The behaviour is fully chained to the stimulus. But how may the system have any relevant active role if its behavioural output is fully chained to the presence of a specific stimulus? The behaviour is wholly determined by the presence of the stimulus. Hence, the conclusion that even though tropistic systems respect the autonomy principle – they are self-moving systems – they do not respect the activity principle – they are fully passive.

So far, so good. On one hand, the complexity, autonomy and activity intuitive principles

are behind some judgments that we have about some representational states; on the other hand, these principles partly *justify* the judgments that some states are representational. Together they constitute the intuitive constraint on proposals for minimal conditions for intentionality. The intuitive judgments and principles are the first two elements playing a role in the application of the reflective equilibrium to the problem of demarcation. What about the third element, that is, the *cognitive theories* that posit representational states to explain behaviour?

These are the theories from cognitive science and other sciences of mind that posit representational states to explain behaviour. There are cognitive theories that posit representational states to explain the behavioural output of rats, vervet monkeys, honeybees, great apes, etc.[60] But how do these theories constitute a constraint on proposals for minimal conditions for intentionality? The posit of a representational state by a cognitive theory is supposed to play a distinctive explanatory role, i.e., it should have an explanatory power that justifies its positing. After all, if the positing of a representational state fails to have any explanatory purchase (and hence the intentional and non-intentional explanations have the same explanatory powers), then the non-intentional explanation should be preferred in virtue of considerations of ontological parsimony. If a given demarcation proposal considers certain states as representational but these states don't play the distinctive explanatory role of representational states in cognitive theories, it follows that this proposal should be rejected. But what is the distinctive explanatory power of representational states in cognitive theories that justifies positing their existence? After all, if explanatory considerations constitute a constraint on proposals for minimal conditions for intentionality, how we are supposed to assess them if we don't know beforehand what is the explanatory power of representational states?

I don't think that a previous specification of the explanatory power of representational states is required for the assessment of demarcation proposals. Indeed, I think that a full

[60] Cf. O'KEEFE & NADEL, 1978; SEYFARTH et al., 1980; VON FRISCH, 1967; CALL & TOMASELLO, 2007.

specification of the explanatory power of representational states goes hand in hand with the assessment of minimal conditions for intentionality. Hence, I adopt the following strategy. I assess the minimal conditions for intentionality established by each demarcation proposal by asking whether the limits of intentionality that it draws is explanatorily justified. Each proposal provides an explanatory justification for establishing minimal conditions. The idea is precisely to assess whether this is the genuine explanatory power of representational states. Does the relevant proposal give rise to a supposed intentional explanation that has an explanatory power different from the non-intentional explanation? Or does it give rise to an explanation that ultimately does not have an explanatory power distinctive from the non-intentional explanation? I conclude that the assessed demarcation proposals are not explanatorily justified. After this negative stage, I develop my own proposal for minimal conditions for intentionality – the *dual proposal* – that is explanatorily justified since it is grounded on what I take to be the genuine explanatory power of representational states.

A consequence of this approach is that it does not follow from the fact that a given successful cognitive theory posits a certain state as a representational that it is genuinely representational. The reason is that there is always the possibility that the relevant state does not play the distinctive explanatory role of representational states and hence its positing as a representation is not explanatorily justified. It is precisely the job of the philosopher of mental representation to deal with this fundamental problem on the nature of intentionality. In fact, cognitive scientists and other scientists of the mind usually do not really assess this problem; rather, they just assume the notion of representational state in the development of the scientific investigation. As Stephen Palmer observes, ‘‘we, as cognitive psychologists, do not really understand our concepts of representation. We propose them, and talk about them, argue about them, and try to obtain evidence in support of them, but we do not understand them in any fundamental sense’’ (PALMER, 1978, p. 259). The philosopher’s job is precisely to provide a

full theoretical understanding of the nature of representational states by making the clarifications and implications of the notion of representationality. That is, to provide a characterization of intentional explanations and how they differ from the non-intentional explanations, a characterization of representational states and how they are distinct from non-representational ones, the specification of the distinctive explanatory purchase of representational states, and so on.

The intuitive conception of representation provides the judgments and principles that constitutes the intuitive constraint on proposals for minimal conditions for intentionality, while the distinctive explanatory role of representational states in cognitive theories constitutes the explanatory constraint on demarcation proposals. But how do these two constraints interact? How may one constraint suppress the other in the context of mutual adjustments that characterizes the reflective equilibrium?

Consider, for example, the scientific discovery of the chemical composition of gold. Before that, people held an intuitive conception of gold – it was whatever intuitively looked like gold, functioned like gold, etc. Such a conception provides paradigmatic cases of gold picked out by our intuitions on gold – e.g., the substance that constitutes certain crowns, medals, and other artefacts, the substance that has certain chemical reactions, etc. These paradigmatic cases of gold are based on human practices and common sense related to gold. However, the development of chemistry led to the discovery that gold is the chemical element 'Au' with an atomic number of 79. That discovery entailed that several substances that previously were intuitively thought to be gold actually were not gold since they did not have an atomic number of 79. Fool's gold (iron pyrite) cases are notorious examples. But why fool's gold is not gold since it looks intuitively like gold? Maybe this is not a scientific discovery about gold, but about another substance since gold just is that substance that for centuries people considered as gold (i.e., both gold and fool's gold).

That objection, however, is implausible. The scientist's job is precisely to specify the chemical composition of the substances that are paradigmatic cases of gold. In the middle of the scientific investigation, it was discovered that the majority of the paradigmatic cases of gold are composed of the chemical element Au, but a minority has a different chemical composition and so are not *genuinely* gold – no matter if they intuitively look like gold. We can make this distinction based on the explanatory role of identifying gold with the element Au in the overall scientific theory. Such discovery has an explanatory significance that justifies ruling out the minority cases as genuinely gold. This identification has the explanatory purchase of unifying all things that are constituted of the element Au as gold. Based on it, science establishes that on one hand, there are substances that we previously thought as not being gold that are gold since they are Au, and on the other hand, there are substances that we previously thought as being gold that are not gold because they are not the element Au. This explanatory criterion constitutes a reason strong enough for trumping the counter-intuitive consequences of the identification of gold with the element Au like ruling out of certain intuitive instances of gold as genuinely gold (fool's gold) or including certain substances as gold that do not intuitively look like it.

The discovery of the composition of gold is an instance of the following scientific investigation. Intuitions pick out the paradigmatic cases of a given substance and further scientific investigation leads to the discovery of the constitutive nature of this substance that entails the revision of some paradigmatic cases that were the starting point of this very investigation as not genuine instances of the substance. That conclusion is based on the explanatory role played by the discovered constitutive nature in the end of the scientific process (in the gold's case, the elemental structure of Au). That is, the scientific investigation of what constitutes these paradigmatic cases leads to the discovery of the constitutive nature shared by a majority but not all paradigmatic cases, which entails the elimination of these minority cases

as genuine instances of the relevant substance. Besides gold, that same scientific process was the case with the discovery of the composition of water (H_2O), the elemental structure of silver (Ag) and other substances.

This process of scientific investigation is analogous to how intuitive and explanatory considerations constitute constraints for demarcation proposals and interact such that one may trump the other (and *vice versa*). The investigation starts with paradigmatic cases of representational states picked out by the intuitive conception of representations (e.g., representational states at the personal level and more simple ones) as well as with intuitive principles on representational states. In parallel, there are paradigmatic cases of representational states posited by successful cognitive theories (e.g., representational states in rats, vervet monkeys, honeybees, etc.). Based on these paradigmatic cases, one engages into the philosophical investigation to specify what is the distinctive explanatory power of representational states. Each proposal for minimal conditions for intentionality specifies what it takes to be the distinctive explanatory role of representational states. It should then be assessed whether it is explanatorily justified, i.e., whether the specified explanatory role is in fact the distinctive explanatory role of representational states. Note that the intuitive principles and the paradigmatic cases are required not only for the identification of the subject matter of the investigation – minimal intentionality – but also for the specification of the explanatory role. After all, if there are no paradigmatic cases, the subject matter of the investigation is up on the air – it may turn out that the investigation is not about representational states anymore. In sum, the paradigmatic cases are the starting point of the investigation of such explanatory role of representational states.

Just like what happened with the case of gold, the specification of the distinctive explanatory role may rule out certain paradigmatic cases as genuine representational states, as well as include other states as genuinely representational. However, different from what

happens on the case of gold, intuitions may rule out certain states as representational even though they play the specified explanatory role. In each demarcation proposal, it should be assessed whether the specified explanatory role of representational states suppress or does not suppress its supposed counter-intuitive consequences. The ultimate goal is that the proposal that results from mutual adjustments between the intuitive and explanatory constraints reaches an optimal equilibrium in which the proposal is defensible in light of both intuitive and explanatory considerations.

The appropriate demarcation proposal is justified by being brought into agreement with the intuitive and explanatory constraints via mutual adjustments. That means that on one hand, an intuitive judgment or principle is rejected if it violates the positing of representational states by a cognitive theory that play an explanatory role that we are unwilling to reject. On the other hand, positing of representational states by a theory should be rejected if it yields assignments of representational status to states that we are not intuitively disposed to accept as genuinely representational. Intuitive judgments, principles and the positing of representational states by cognitive theories are revised via mutual judgments, going back and forth on what these constraints dictate until they are harmonious.

We can illustrate such mutual adjustments in the following ways. Suppose that we have a certain intuitive judgment that a given state is not genuinely representational. Nevertheless, on one hand, this state is posited by a successful cognitive theory and it plays the distinctive explanatory role; on the other hand, it respects several intuitive principles and the intuitive judgment that it violates is not highly entrenched in our intuitive conception of representation. It is plausible to adjust the intuitive constraint by setting aside this intuition. That is, the non-intuitive aspects of a posited representational state should be sufficiently compensated by the explanatory role of positing this state. If so, it becomes plausible that it is a genuine representation. Now suppose that there is a cognitive theory that posits a certain state as a

representation. Nevertheless, on the one hand, this state does not respect several intuitions that are highly entrenched in our intuitive conception of representation; on the other hand, the positing of this state as a representation does not play an indispensable explanatory role. Therefore, it is plausible to reject the assignment of a representational status to this state.

It is expected that the appropriate demarcation proposal is the one vindicated by these intuitive and explanatory constraints in the end of this process of mutual adjustments. That will be the optimal state: there is an optimal equilibrium between the intuitive principles and the explanatory role of representational states posited by cognitive sciences. But how can one be assured that this optimal state is achievable after all? The fact is that there is no assurance beforehand. Only in the end of this process is that one can give a secure response to this question.

The method of reflective equilibrium provides a plausible way to deal with the second methodological problem on the status of intuitions: is the intuitive conception of representation inconsistent? Inconsistency is always a real threat when dealing with a conception of a given phenomenon which has not passed through the sieve of a rigorous theoretical investigation on the nature of this phenomenon. That is the case of intuitive conceptions in general and of the intuitive conception of representation in particular. In fact, there are several candidates of inconsistent intuitive views. As an illustration, I will briefly describe what some take to be an inconsistency on our intuitive conception of free will.

Galen Strawson and others have argued that there is an inconsistency between our intuitions on free will. A free action is an action for which the agent is truly responsible for it. Determinism is the view that every event (in conjunction with the laws of nature) is causally necessitated by antecedent events. We have an intuitive conception of free will according to which free actions are not determinate and not random. We cannot have free actions if our actions are wholly determined by causes anterior to our existence. But if our actions are not

determined, we cannot have free actions either. That is the case because if our actions are not determined, then they would be totally or partly random, but we cannot be responsible for our actions if they are random. The conclusion is that free will is incompatible with both determinism and indeterminism: there is an inconsistency between the intuition that free actions are not determinate and the intuition that free actions are not random.[61]

However, even assuming that there are potential incompatibilities among our intuitions on the limits of intentionality, it does not follow that they cannot constitute a genuine constraint on demarcation proposals. The reason is that these potential inconsistencies may be fixed by the mutual adjustments between intuitive and explanatory constraints that characterize the method of reflective equilibrium. In this revision process, inconsistent intuitions may be fixed via the rejection or modification of one or more intuitions. Alternatively, in the end of the theoretical investigation of the limits of intentionality, we may find out that the relevant intuitions are not incompatible at all. The worry that they are inconsistent is explained away. Thus, the concern that the intuitive conception of representation is inconsistent is not a reason to throw away the intuitive constraint on proposals for the limits of intentionality.[62]

Let me finally describe my strategy of adopting the method of reflective equilibrium in order to develop my proposal for minimal conditions for intentionality. It is divided into two stages. The first one is negative: it is the assessment and rejection of two demarcation proposals, the *causal independence proposal* in the next section (FODOR, 1986; BECKERMANN, 1988) and the *constancy mechanism proposal* (BURGE, 2010, pp. 292-308; STERELNY, 1995; RESCORLA, 2013). I assess them in light of both the explanatory and

[61] Cf. STRAWSON, 1986, p. 25-6. This argument, however, is contentious. There is a whole literature on free will and determinism and It is not my goal to assess this argument here. I have just presented it to illustrate a candidate of inconsistent intuitive conception of a given phenomenon.

[62] But what if our intuitions are so utterly inconsistent that they cannot be fixed? The answer to this question cannot be given in advance. Only in light of the possible inconsistent intuitions is that this question is answerable. But the fact that they may be fixed via mutual adjustments is enough to show that possible inconsistencies are not a reason for an a priori rejection of intuitions as a genuine constraint on demarcation proposals.

intuitive constraints, but primarily focused on the explanatory constraint. I have previously defended the autonomy, complexity and activity principles as the most entrenched principles in our intuitive conception of representation. But what is the nature of the distinctive explanatory role that constitutes the explanatory constraint? This is a hard problem that will occupy us in much of what follows.

Fodor and Beckermann propose that it is a minimal condition for intentionality that the relevant state is causally independent of the external condition which it supposedly represents. This demarcation proposal gives rise to the objection that teleosemantics and many other naturalist theories are too liberal since they treat certain states that are not causally independent of the relevant external condition as genuine representations. By contrast, Burge, Sterelny and Rescorla objects that teleosemantics and many other naturalist theories are too liberal by appealing to the constancy mechanism proposal. According to this proposal, it is a minimal condition for intentionality that a system employs constancy mechanisms in the production of representational states. Since teleosemantics and many other naturalist theories treat certain systems that do not employ such constancy mechanism as genuinely representational, it follows that these theories are too liberal. I argue that these proposals fail to specify the distinctive explanatory role of representational states and hence that they are not explanatorily justified. Finally, they are not intuitive on relevant aspects and since there is no explanatory argument to compensate it, the conclusion is that these proposals should be rejected.

The next chapter is wholly dedicated to the positive stage. There I develop my own proposal for minimal conditions for intentionality – the *dual proposal*. I start by developing the success pattern proposal. However, this proposal is flawed because the success pattern condition is too liberal on the requirements for positing of representational states to be explanatorily justified. It is also too liberal in light of the intuitive principles for a state to be representational. Therefore, it is necessary to revise the success pattern proposal as dictated by

the mutual adjustments that characterize the method of reflective equilibrium. I revise it by adding another minimal condition for intentionality – the *constancy mechanism condition*. The result is the dual proposal for the limits of intentionality that establishes two minimal conditions – the success pattern and the constancy mechanism conditions. I then argue that the dual proposal establishes the distinctive explanatory role of representational states and so is explanatorily justified. Furthermore, it is immune to the objections that I raise to the causal independence and the constancy mechanism proposals, even though it establishes a variation of the constancy mechanism condition that was proposed by the constancy mechanism proposal. Finally, I argue that the dual proposal vindicates the intuitive conception of representation. The result will be the optimal state in the revisionary process of the method of reflective equilibrium. The mutual adjustments between the explanatory role of positing representations and the intuitive conception of representation reaches the optimal equilibrium in which we are not inclined to revise the resulting proposal anymore, i.e., the dual proposal.

3.4 The causal independence proposal

Jerry Fodor and Ansgar Beckermann have proposed that a given state is not a genuine representation unless there is a causal independence between the external stimulus which triggers the tokening of the state and the output behaviour triggered by the state. The criterion is that if there is a causal dependency between the tokening of the relevant stimulus and the tokening of the supposed representational state (and hence with the output behaviour), then this is not a genuine representational state. But if there is no such causal dependence, then it becomes plausible to attribute a representational status to the relevant state.[63] That is the causal independence proposal (FODOR, 1986; BECKERMANN, 1988).

Consider tropistic systems like amoeba and paramecia. Are paramecia genuinely

[63] Note that this is not a sufficient condition for intentionality, but rather a minimal condition for intentionality.

representing the presence of light when they have an avoidance behaviour towards it? Or they are just indicating or registering, not genuinely representing the presence of light? Generalising the question, do tropistic systems really represent the relevant external stimuli that triggers the token of the supposed representational states and thus the subsequent responsive behaviours?[64]

According to the causal independence proposal, tropistic systems like paramecia are not genuine intentional systems. The problem is that the there is a causal dependence between the presence of the external stimulus and the triggering of the internal state. That is, there is a causal relation which establishes that the presence of the external stimulus is the cause of the tokening of the state and so of the production of the relevant behaviour. For instance, there is a causal relation between the presence of light in the surrounding environment of the paramecium and the tokening of the internal state and the avoidance behaviour. Hence, the paramecium's internal state is not a genuine representation of light. This is a deep contrast with genuine representational systems in which the relation between the relevant external stimulus and the tokening of the representational state is not causal. The causal independence proposal claims that a given sensory state is genuinely representational only if there is a causal independence between the relevant stimulus and the tokening of the state (and hence with the response behaviour). That is, it is possible in a given situation to have the external stimulus but not the relevant state, since they are not causally dependent.

But what is the nature of this causal dependence? Here Fodor and Beckermann diverge. Fodor claims that it is a lawful relation – a natural law which establishes that the presence of the relevant external condition causes the tokening of the internal state and hence the triggering of the response behaviour. There is a lawful covariation between the presence of the external condition which constitutes the external stimulus and the production of the sensory state which ultimately triggers the output behaviour. In Fodor's words, "any system that can respond

[64] Here I am just conceding, for the sake of the argument, that tropistic systems have genuine behaviour.

selectively to nonnomic properties is, intuitively speaking, a plausible candidate for the ascription of mental representations; and any system that can't, isn't" (FODOR, 1986, p. 11). Thus, the fundamental distinction is that non-intentional systems respond to properties of the relevant external condition in a nomological way, while intentional systems respond to properties of the external condition in a non-nomological way.[65]

In contrast, Beckermann rejects this non-nomic requirement and maintains that for a system to be intentional it is required that there is no causal dependence, regardless whether there is a natural law supporting it or not.[66] That is, the system is intentional provided that there is no causal dependence between the tokening of the internal state and the relevant external condition, no matter whether such dependence is nomological or not.[67] Hence, Beckermann and Fodor provide different formulations of the causal independence proposal. In what follows, I put aside this controversy and assess the general formulation of the proposal according to which a state is representational only if there is no causal dependence between its tokening and the presence of the relevant stimulus. After all, if it turns out that causal independence is not a minimal condition for intentionality, it makes no difference whether the relevant causal dependence relation is nomological or not.[68]

What are the arguments supporting the causal dependence proposal? Consider a situation in which there is a covariation between the response behaviour of a system and the presence of a given external condition. There is no causal dependence between the response

[65] Fodor has since changed his position, but the extent of this change is not clear, cf. FODOR, 1991, p. 257.
[66] "In my eyes, however, that is [Fodor's non-nomic requirement] asking too much. For, as I see it, the basic criterion is not that a behaviour which constitutes a selective response to a specific feature of the environment *cannot* be caused by this feature but that it is in fact not caused by it." (BECKERMANN, 1988, p. 140).
[67] Fodor's nomological requirement is susceptible to counter-examples. For instance, Kim Sterelny has noted that desert isopods can distinguish their relatives through chemical cues and being a relative of isopod number 47,012 is not a nomic property. But it is highly implausible to attribute intentionality to desert isopods (cf. STERELNY, 1995, p. 251 - 270). Moreover, Louise Antony and Joseph Levine have pointed out that Fodor's distinction of nomic and nonnomic properties is problematic (cf. ANTONY & LEVINE, 1991, pp. 3-7).
[68] This general formulation of the proposal is not committed to the thesis that causes necessitate their effects. So, for any two events *x* and *y*, there is at least one situation in which *x* is the case, but not *y*.

behaviour and the external state, that is, the external condition does not cause the responsive behaviour. But if there is no causal dependence, why is there such covariation? What explains the fact that there is a covariation between the response behaviour and the external condition given that there is no causal relation at all? Notice that there is no third state which causes the presence of the external condition and the responsive behaviour.

The explanation of this fact, argues Fodor and Beckermann, is not a non-intentional explanation, but an intentional one. The best available explanation consists in positing a representational state which represents the external condition and triggers the response behaviour. There is a covariance between the external condition and the response behaviour because the system has a state which represents the external condition and the tokening of the state triggers the response behaviour. But the relation between the system's state and the external condition is not causal – the external condition does not cause the tokening of the state. Rather, there is a *semantic* relation between them which explains why there is a covariation even though not a causal relation. Therefore, the semantic relation between the system's state and the external condition is what explains this covariation. Namely, the covariation between (i) the external condition and (ii) the tokening of the system's state and the subsequent output behaviour. The explanation of the mystery of the non-causal covariation between the external state and the response is "the great evolutionary problem that mental representation was invited to solve" (FODOR, 1986, p. 14).

But what is this semantic relation? How is it possible? Fodor and Beckermann argue that the intentional system infers from the presence of a certain property (or properties) of the environment which is causally related to the system, the presence of other environmental property that is not causally related to the system. That relation between the system's state and the non-causally related property of the environment constitutes the semantic relation between them, with the system's state representing the property. The inference is what Fodor calls a

"selection" and Beckermann calls a "choice" (FODOR, 1986, p. 11; BECKERMANN, p. 132).

The idea is that the intentional system makes a perceptual inference from the presence of a certain property (or properties) in the environment to which the system is causally related, to the presence of another environmental property to which the system is not causally related. That latter property, not the former, is the one which is represented by the system. The perceptual inference is responsible for the tokening of the representational state and ultimately for the triggering of the behavioural response. But how does this inference occur? Typically (but not exclusively) the inference occurs via a covariation between the causally related and the non-causally related properties of environment in the sense that the indication of the presence of the first property leads the system to infer the presence of the second property.[69]

In sum, the explanation of the covariance between the external condition and the responsive behaviour despite the absence of a causal relation between them is the existence of the system's representational state. The representation is not causally related to the external condition but nevertheless represents it via an appropriate inference which ultimately triggers the responsive behaviour. The inference bridges the gap between the properties of the environment that are causally related to the system[70] and the properties of the environment that are not causally related to the system. Hence, the positing of an internal representational state is explanatory justified because it plays a crucial and indispensable explanatory role in the explanation of the covariance between the external state and the response behaviour of the system.

But when is the positing of a representational state not explanatorily justified? Precisely

[69] Considering the represented property *O* of an external object *S* by a given intentional system, "presumably such [perceptual] inferences exploit information from memory as well as information about the detected [...] properties of *S*; in the typical [..] case, this would be information to the effect that the [detected] properties cohabit with property *O*, so that the detection of the former provides a reliable index of the presence of the latter" (FODOR, 1986, p. 14). This is the so called "perceptual inference" of classical intentional psychology.

[70] Since we are talking about the relevant stimulus to which a tropistic system is causally sensible, the properties of the external condition to which tropistic systems are causally related are not distal, but proximal properties.

when there is a non-intentional relation between the response behaviour of a system and a given external condition. In this case, the explanation of the covariance is clearly causal – the presence of the external state causes the response behaviour. That is what happens in tropistic systems – there is a causal connection between the response behaviour of the tropistic system and the presence of the external condition. Since the non-intentional explanation fully explains the behaviour of the tropistic system – nothing is left to be explained – then the appeal to any intentional notion in order to explain the system's behaviour is not required. So, paramecia have avoidance behaviour when there is an appropriate light in the surrounding environment because the light directly causes the avoidance behaviour. That is a non-intentional explanation, it does not appeal to any intentional notion. Hence, the positing of the representational state in cases where there is a causal connection between the external condition and the system's response behaviour is dispensable. It does not satisfy the explanatory requirement and so fails to earn its explanatory keep. In light of the ontological parsimony requirement, given that intentional and non-intentional explanations of the system have the same explanatory power, it follows that the non-intentional one should be preferred.

The causal independence proposal seems prima facie to be plausible. It draws a distinctive and clear-cut line between genuine representational systems and tropistic systems by appealing to a plausible criterion of causal independence that constitutes the explanatory role of representational states. What could be wrong with it? Before assessing this proposal, it is necessary to make the following elucidation.

The problem whether tropistic systems are genuinely intentional may be characterised as the question whether the system's state constitutes a representation of the external stimulus. It assumes that there is an intermediate internal state between the presence of the relevant stimulus and the tropistic system's behavioural output and asks whether such state is representational or not. However, this characterization leaves out a third possibility – maybe

there is no intermediate state at all; the presence of the relevant stimulus directly causes the behavioural output. Hence, the problem may also be characterised as the question whether there is an intermediate internal state, and if so, whether it represents or does not represent the relevant stimulus. Here there is no assumption of an intermediate internal state and the door is open for three different responses: (i) there is an intermediate internal state which represents the relevant stimulus; (ii) there is an intermediate internal state but it does not represent the relevant stimulus (maybe it is just a sensory state that registers or indicates such stimulus); (iii) there is no intermediate internal state. In what follows, I do not commit myself to either the assumption that there is an intermediate internal state (representational or not), or with the assumption that there is no intermediate internal state. Thus, I characterise the problem as whether tropistic systems are intentional or not. In case in which they are not intentional, there is still the possibilities that there is or not an intermediate internal state.[71]

That said, let us assess the core claim of the causal independence proposal. The system's state represents a causally independent external condition because (i) the system infers its presence from the presence of a second external condition; and (ii) there is a covariation between these conditions and this second condition is causally related to the system. But if so, why not claim that a system's state that is caused by an external condition which by its turn is caused by a second external condition represents this second condition? That is, why not claim that the system's state represents the condition in the end of this causal chain, namely, the second external condition? This case, however, is easily ruled out by the causal independence proposal. If the system's state I is caused by external condition A which is caused by external condition B, then I is caused by B and so there is a causal dependence that rules out I as a representation of B. So, the proposal rules out the intentionality of *direct*

[71] Notice that the fact that there is no intermediate sensory state is fully compatible with the fact that there are other intermediate internal states that trigger behaviour but are neither representational nor sensorial.

causal dependence (i.e., X is not a representation of Y because X is caused by Y) and also of *indirect* causal dependence (X is not a representation of Z because X is caused by Y that is caused by Z).

So far, so good. Now suppose that in a given situation I and B are both tokened but not A. For instance, a situation in which I and B are tokened by another external condition C that independently causes I and B but does not cause A. So, in this case there is no indirect causal relation between I and B. That is, there is no single causal chain connecting I and B. Rather, there are two distinct causal chains responsible for their tokens. Now notice that an indirect causal dependence is a weaker causal dependence relation than a direct one. Finally, one can weaken the indirect causal relation even further by stipulating that the causal chain is not constituted by only three states, but by four, five states, etc. But then it becomes implausible to claim that intentionality is ruled out in indirect causal dependence cases but not in the above non-causal covariation scenario. If even such weaker causal dependence rules out intentionality, why does the covariation relation not also rule out intentionality? What is the fundamental difference between them? It is implausible to say that this is so because the weaker causal dependence is a stronger relation than the covariation relation since the possibility of a perfect covariation relation is not excluded. So, why does the weaker causal dependence rule out intentionality, but a perfect covariation relation does not rule it out? It is not a good reason to claim that this is the case because the weaker causal dependence is still a stronger relation than a perfect correlation. I think that this is problematic for the causal independence proposal, but it does not constitute a knock-down objection. A stronger objection is required to demonstrate that this proposal is not viable. In what follows, I present what I take to be the fundamental problem with it.

Fodor and Beckerman are right in claiming that in cases where the behaviour of the relevant system is fully explainable in non-intentional terms, the positing of the system as

intentional is dispensable. There is nothing missing in the explanation of the system's behaviour that could justify the assignment of a representational state to it. However, what is crucial for the conclusion that tropistic systems are non-intentional is not that the relevant states are causally dependent on the stimuli that trigger their tokens. Rather, it is that positing these states as representations plays no distinctive role in the explanation of tropistic behavioural outputs.

The fundamental distinction between tropistic and genuine intentional systems does not rest on the fact that genuine representations are causally independent from the properties of the external environment which they represent, but rather on the fact that the positing of some internal state as a representation plays a distinctive role in the explanation of the system's behaviour. To illustrate this view, consider a mechanical system which drains the water of a bathtub whenever it is full. In order to do it, the drain system has a sensor which detects the level of water in the bathtub and thus whenever the sensor indicates that the bathtub is full, the system starts to drain the water. Therefore, the stimulus which activates the drain system is a certain amount of water in the bathtub.[72] Does the system really represent this stimulus?

It is absurd to defend this conclusion precisely because you can fully explain how the drain system works without positing that it is an intentional system which represents the amount of water inside the bathtub. The attribution of intentionality to the drain system plays no explanatory role in the explanation of how this system works. The case of tropistic systems is analogous: in both cases, the fundamental reason that the systems are not genuinely intentional is not that the internal states are causally dependent on the relevant stimulus, but

[72] Alternatively, in order to avoid the commitment with an intermediate state between the stimulus and the drain of water which constitutes the sensor of water, one may just suppose that the bathtub has a hole with the appropriate size to prevent the leaking of water. In this case, the state of the bathtub being full of water would directly cause pouring of water via this hole.

rather that the positing of a representational state is explanatory idle.[73] The non-intentional explanation fully explains their behavioural outputs, nothing is left to be explained by the intentional explanation. *That* is the fundamental reason that tropistic systems are not genuinely intentional.

Finally, causal dependence is only one of the reasons that positing a given system as representational does not play the explanatory role. Notice that there are situations in which the positing of an internal state as a representation plays no explanatory role even though it is causally independent of the stimulus which triggers it. In addition to causal dependence, there are other reasons for the positing of an internal state as representational to be explanatory idle.

Let's consider the tokening of a system's internal state that covaries with a given external condition, but nevertheless the presence of this external condition does not cause the tokening of the state. Now suppose that there is a second external condition that causes both the first external condition and the tokening of the system's internal state. Hence, there is no causal dependence between the internal state and the first external condition. Furthermore, by coincidence, the tokening of the internal state makes the system to move in the direction of the first external condition. So, does the internal state represents the first external condition? Isn't the approaching behaviour of the system towards the first external condition a strong evidence that the system represents it? Not at all. The system's behaviour is fully explainable by the non-intentional explanation. What happens is that the non-intentional explanation that appeals to the second external condition to explain the system's behaviour has the same explanatory power as the intentional explanation that posits the system as representing the first external condition. Therefore, the conclusion is that positing the internal state as a representation of the first external condition is explanatory idle, even though there is no causal dependence between

[73] Evidently, both cases are highly distinct in relation to the question whether intuitively they are intentional systems or not. But here I am focusing on the explanatory aspect of positing that these systems are intentional. Only later I will assess the intuitiveness of tropistic states as representational.

them.

This is only one example of a situation in which the explanatory requirement for positing an internal state as representational is not satisfied for a reason other than causal dependence. In this case, the relevant reason is that there is a mere covariation of the internal state and the first external condition in virtue of a second external condition, which causes the tokening of both in a complete independent way.[74] Causal dependence is just one of the reasons that the positing of a system's internal state as a representation may be explanatorily idle and so the ultimate and fundamental distinction of representational states from tropistic and similar states does not lie on the distinction between causal dependence and independence, but rather on the distinction between the explanatory idleness and the explanatory power of positing of internal state as representational. Contra Fodor and Beckerman, that is the condition for minimal intentionality that one should keep an eye on.

But there is a problem behind this verdict that positing certain systems as intentional is explanatorily idle which was still not properly answered. What is the precise nature of a justified appeal to an intentional explanation of the behaviour of a given system? That is, what is the distinctive explanatory role that positing a representational state should play for it to be explanatorily justified? It could be objected that only in light of the specification of such distinctive explanatory role is that it is possible to conclude that positing certain systems as intentional is not explanatorily justified. So, it is not possible to determine that positing tropistic systems as intentional systems is explanatorily idle. However, this objection is not viable.

The previous specification of the distinctive explanatory role of representational states

[74] It could be replied that in this example the causal dependence between the internal state and the second external condition is the reason that positing the internal state as representational is explanatory idle. That is true but note that the causal independence proposal claims that a state is not a representation of an external condition provided that the former is causally dependent on the latter. This criterion does not preclude in any way whatsoever the internal state to represent the first external condition when there is causal dependence between this state and a second external condition (not between the state and the first external condition), as it happens in the above example. So, this is a reason for the explanatory idleness of positing a representational state different from the causal dependence reason.

is not required to conclude that positing certain systems as intentional is not explanatorily justified or that a given proposal fails to specify the distinctive explanatory power of representational states. The verdict that the causal independence proposal fails to specify such explanatory role is not precluded in the absence of such specification for two reasons. First, this verdict is based on counter-examples to the causal independence proposal – there are systems that although they satisfy the causal independence condition, there is no explanatory justification for their positing as intentional systems. Second, this verdict is also based on the fact that the causal independence proposal fails to specify what the non-intentional explanation leaves to be explained by the intentional explanation in cases in which the positing of representational states is explanatorily justified. In conclusion, the causal independence proposal fails to specify the explanatory role of representational states. But what is left to be explained? Above I have just assumed that there is something left to be explained, otherwise there would be no distinctive explanatory power of representational states in intentional explanations. The goal of the ongoing investigation is precisely to specify what is this explanatory power. Finally, this negative stage of ruling out proposals that fail to specify the explanatory role of representational states is required for the further development of this investigation. As it will be clear in the next chapter when I will specify this explanatory role, this negative stage will help to reach the positive stage in which this explanatory power is finally specified.

What about the intuitiveness of the causal independence proposal? Is it compatible with the intuitive conception of representation? It is clear that positing tropistic systems as intentional systems violates principles of our intuitive conception of representational states. It is true that the causal independence proposal rules out tropistic systems and other very simple systems as genuinely intentional since they fail to satisfy the causal independence condition. Hence, it seems that this proposal is compatible with the complexity principle which establishes

that systems without a certain cognitive complexity are not genuinely intentional. However, there is no guarantee that the satisfaction of the causal independence condition by a given system entails that it has a minimal cognitive complexity. The previous example of the internal state that covaries with the first external condition in virtue of the presence of a second external condition which independently causes the tokening of both satisfies the causal independence proposal. However, there is no guarantee that such system has a minimal cognitive complexity. Since the system is causally chained to this second external condition and automatically moves into the direction of the first external condition only in virtue of this causation, it is deprived of a minimum cognitive complexity. Evidently, it is an open question what is the extent of such minimal complexity, but it is hard to argue that the mere fact that there is no causal dependence between the relevant state and the external stimulus is enough for the system to respect the complexity principle. The required cognitive complexity is more demanding than that.

What about the autonomy principle? It requires that the forces responsible for the system's behavioural output originates within the system, not outside. The causal independence condition rules out several systems that approach or avoid certain external stimulus in virtue of forces that originate outward the system because in these cases there is a causal dependence between the system and the relevant external stimulus. So, it is plausible to claim that the causal independence proposal does not violate the autonomy principle.

Finally, there is the activity principle which requires that the system cannot be wholly passive for it to be intentional. Some active role is required. The causal independence condition rules out the needle as an intentional system in which the forces that causes its movement in the direction of the magnet comes only from the magnet. That is the case because here there is a causal relation between the system and the magnet. But the causal dependence condition fails to rule out intentionality in the aforementioned example in which the internal state covaries with the first external condition by being caused by a second external condition. In this case,

the system's behavioural output towards the first external conditions is automatic, it is completely determined by the presence of this second external condition and hence there is no space for the system to have any active role. Thus, the causal independence proposal violates the activity principle.

In the context of mutual adjustments between the theoretical virtues of positing representational states and the intuitive conception of representation, it is required a strong explanatory reason for the causal independence proposal to override the complexity and activity principles. That is, a strong explanatory reason, that together with the compatibility of the causal independence proposal with the autonomy principle, justifies the acceptance of this proposal despite its violation of the complexity and activity principles. Is such explanatory reason available? As previously showed, the causal independence proposal is not explanatorily justified, it fails to specify the distinctive explanatory power of representational states. Such explanatory reason is simply absent.

In conclusion, the causal independence proposal is flawed because it fails to specify the distinctive explanatory power of representational states and thus it is not explanatorily justified. Furthermore, it is not intuitive since it violates the complexity and activity principles that are highly entrenched in our intuitive conception of representation. That said, let us move on and assess the constancy mechanism proposal.

3.5 The constancy mechanism proposal

Some philosophers claim that it is plausible that a given system represents a certain distal feature but not the proximal stimulus because even though at different instants there is a great variety of proximal stimuli reaching the system's sensory apparatus, the distal feature and the system's response behaviour remains the same. Hence, the conclusion that the same thing is being represented throughout all these changes in proximal stimuli, namely, the distal object.

That is the line of reasoning behind this conclusion. Given that there is a great variety of proximal stimuli coming from the distal object and still the distal object and the system's responsive behaviour remains the same despite all varieties in proximal stimuli, it follows that the system is representing the distal object, not the proximal stimuli.

In order to achieve this result, the system should employ a constancy mechanism, that is, a mechanism that guarantees that the system still represents the distal feature despite huge varieties (to a certain extent) in proximal stimuli coming from the environment. There are several examples of constancy mechanisms: *colour constancy* is possibly the most famous. A given visual system sees an object as having the same colour even when there are huge differences in the environmental light conditions. The system keeps the representation of the object's colour constant despite a great variety of light reflected by the object under different lighting conditions and so a great variety of light rays reaching the retina. To illustrate colour constancy, consider the following cases. There is a white cup which appears to us as having a uniform colour under a highly uneven illumination, despite the fact that the light reflected by the cup's shaded region is very different from the light reflected by the unshaded one. This is a case of colour constancy – the visual system represents all regions of the cup as having the same colour even though there is a huge variation in the illumination incident on them. Now consider a case of *shape constancy*: a coin looks round when viewed head-on as well as when viewed from acute angles, despite the areas projected by the coin on the retina are hugely different under these two conditions. This is a case of shape constancy because the coin looks as having the same shape even though it is seen from very different angles and so there is a huge variety in the light intensities reflected by the coin. Other cases of constancy mechanisms include size constancy, position constancy, etc.[75]

[75] This is a very simplified characterization of constancy mechanisms. A proper characterization would have to assess different problems, like the ones on similarities and dissimilarities aspects of the appearance of the object throughout changes of proximal stimuli (e.g., colour, shape, etc.). It is not my goal to enter into this debate here, for a discussion on the proper characterization of constancy mechanisms, cf. COHEN, 2015; HILBERT, 2005.

In light of this feature of constancy mechanisms, several philosophers have defended the employment of the constancy mechanism as a minimal condition for intentionality. Tyler Burge has defended that the limits of perception are the limits of intentionality and that what distinguishes perceptual and non-perceptual states is that perceptual states employ constancy mechanisms.[76] Hence, what distinguishes representational states from non-representational states is that the first ones employ constancy mechanisms. According to Burge, "certain processes in perceptual systems systematically distinguish effects of stimulation that are special to the individual and the context from perspective-independent attributes of the wider environment. Explanation of the formation of perception keys on processes in perceptual systems that make this distinction. Such processes constitute the ground of perception, representation, and objectivity" (BURGE, 2010, p. 23). Kim Sterelny also reaches the same conclusion, even though he is not committed to Burge's thesis that the limits of intentionality are the limits of perception. He defends that it is a minimal condition for a given state to constitute a representation that "there is a sufficient variety of proximal routes and sufficient stability of distal sources" (STERELNY, 1995, pp. 261-2).

It strikes me that one of the motivations behind the constancy mechanism proposal is the following. Philosophers like Fred Dretske have defended that the employment of constancy mechanisms is a condition for the system to represent the distal feature, not the proximal stimuli.[77] That is, the appropriate criterion to determine when an intentional system represents the distal feature, not the proximal stimuli that reach the sensory apparatus, is when the system employs a constancy mechanism in the production of the representational state. The seeming initial plausibility of the constancy mechanism criterion leads to the temptation of using it not

[76] "[...] perception marks the lower border of representation. Perception lies not only at the root of empirical objectivity. It is, I think, where states with veridicality conditions first clearly emerge." (BURGE, 2010, p. 549).
[77] Dretske is famous for being one of the first philosophers for appealing to the system's employment of constancy mechanisms as a criterion for determining whether the system is representing the distal object or the proximal stimulus, cf. DRETSKE, 1981, p. 163.

only to determine when the system is representing distal objects and not proximal stimuli, but also to demarcate the limits of intentionality.[78] According to the resultant demarcation proposal, it is a minimal condition for a system to be intentional that it employs the constancy mechanism. However, this move is a *non sequitur*.

It is a great leap to infer from the thesis that a system represents a distal feature in virtue of the employment of the constancy mechanism to the thesis that the employment of the constancy mechanism is a minimal condition for intentionality. First, the constancy mechanism criterion implicitly assumes that the relevant system is an intentional system – after all, to ask whether the system represents the distal feature or the proximal stimuli is already to assume that the relevant system is intentional. Second, it is plainly possible that a non-intentional sensory system produces a state that correlates with a distal feature of the external environment even when the proximal stimulus reaching its sensory apparatus varies a lot. That may happen because there is a third state which causes both the production of the system's sensory state and the occurrence of the distal feature.[79] Finally, it is interesting to note that Dretske, famous for holding the constancy mechanisms criterion for determining distal content, refuses to hold that the employment of constancy mechanisms is a minimal condition for intentionality – i.e., that it is required for a system to be intentional in the first place (DRESTKE, 1981, p. 163; 1986, p. 168-171).

Since the thesis that the system represents distal objects in virtue of the employment of constancy mechanism does not imply the thesis that such employment is a minimal condition for intentionality, what other arguments are there for the constancy mechanism proposal? What

[78] Here I am not committing myself with the plausibility of the claim that the employment of the constancy mechanism is a condition for a system to represent a distal feature. I am just describing what I take to be one of the main motivations behind the constancy mechanism demarcation proposal in order to assess it. I will effectively assess the distal content problem for the teleosemantic account of content only in the fifth chapter.
[79] Suppose that you have a system with an internal state that merely covaries with a certain external state because there is a second external state which causes both the internal state and the first external state. The proximal stimuli may change a lot and nevertheless the internal state and the first external state will continue to covary as long as the second external state causes the presence of both states.

other reasons are there for the thesis that the employment of constancy mechanism draws the limits of intentionality?

Kim Sterelny and, followed by him, Peter Schulte have proposed the counterfactual robustness argument in favour of the constancy mechanism proposal that appeals to a distinction between two kinds of explanations, namely, *robust-process* and *actual-sequence explanations* (STERELNY, 1995, p. 258-262; SCHULTE, 2015). This distinction is well illustrated via the explanation of the result of a football match. Consider the final of the 1970 FIFA World Cup in which Brazil defeated Italy. You may explain the result of this football match in two different ways. On one hand, there is the actual-sequence explanation which consists in a detailed description of the match, describing every pass, free kick, cross, goal, shooting, and so on – that is, a complete physical description of every move in the match. On the other hand, there is the robust-process explanation which consists in a description of the abilities of Pelé, Rivellino, Tostão and other Brazilian players and concludes that given their superior quality over the Italian players, Brazil would defeat Italy in one way or another. As it happens, one explanation has advantages over the other, the reason that they are not replaceable. The actual-sequence explanation is more specific (i.e., it has a complete description of everything that happened in the match, including a description of why Brazil won by 4x1), but the robust-process explanation has the advantage of explaining what would probably have happened if things had been a little bit different (e.g., Brazil, not Italy, had the kick-off) – Brazil would still probably win.

The lesson that Sterelny has drawn from the robust-process and actual-sequence explanations distinction is that intentional explanations are robust-process explanations, not actual-sequence ones. An actual-sequence explanation of the behaviour of a given organism describes the precise sequence of neurological and physical events which lead to the behaviour, while an intentional explanation of it is a robust-process explanation. But what is so special

about the intentional explanation as a robust-process explanation in such a way that the actual-sequence explanation of behaviour misses? What is left behind? The case here is analogous to what happens with the robust-process explanation of the result of the football match – the intentional explanation of behaviour explains how the organism would have behaved if things were a little bit different.

Let's contrast the actual-sequence and robust-process explanations of the avoidance-behaviour of an animal when it stares at something in his visual field.[80] The actual-sequence explanation specifies the shadow in the animal's retina, every detail of what happened in his brain, limbs, etc. However, it cannot explain what would have happened if the animal were in a slightly different position, or if the perceived object were in a slightly different angle, etc. In contrast, the intentional explanation explains that the animal had this avoidance-behaviour because it saw a predator. For instance, in case of the animal were not in position *x* but in position *y* or in case of the predator were not in the angle α but β, the animal would still have the behaviour of trying to avoid it. In sum, intentional explanations of behaviour are capable of giving an account of counterfactual scenarios, while actual-sequence explanations are not.

As it happened with the robust-process explanation of the result of football match, the intentional explanation of behaviour has the advantage of explaining what would have happened in different situations. It loses in richness of detail in order to gain in systematicity. While actual-sequence explanations of behaviour appeal to proximal features (e.g., a shadow in the retina with a certain form), intentional explanations appeal to distal features of the environment which are represented by the organism (e.g., the representation of a predator nearby which may be triggered by shadows of slightly different forms in retina). The intentional explanation explains behaviour as a response to a distal feature of the environment (e.g., the

[80] As an actual example of it, Sterelny mentions the piping plover avoidance behaviour when it stares a predator approaching its nest, which consists in the feigning of a broken wing in order to seduce the predator into its direction and hence to keep it away from the nest (RISTAU, 1991; STERELNY, 1995, p. 261).

representation of a predator), which may be triggered via widely different proximate stimuli (e.g., different shadows in the retina).

In light of this characterization of the distinction between intentional and non-intentional explanations, it becomes clear why only systems that employ constancy mechanisms have behavioural patterns that are prone to intentional explanations. Constancy mechanisms allow the system's behaviour to be triggered by the same distal feature despite a great variety of proximal stimuli. Without the employment of constancy mechanisms, the system's behaviour is triggered only by proximal stimuli. Since intentional explanations are robust-process explanations, the intentional explanation of the system's behaviour appeals to the system's response to distal features of the external environment and the only way that the system can respond to distal features but not to proximal stimulus is via the employment of constancy mechanisms. Hence, the conclusion that the employment of constancy mechanisms is a condition for the system's responsive behaviour to be properly explained in intentional terms and so that constancy mechanisms constitutes a condition for minimal intentionality.[81] What could be problematic with the counterfactual robustness argument?

One could complain that it is not clear why a distinction between robust-process and actual-sequence explanations in fact consists in a distinction between intentional and non-intentional explanations. That is, that the distinction between robust-process and actual-sequence explanation do not demarcate a line between intentional and non-intentional systems, but merely a line between systems that employ constancy mechanisms and systems that do not employ. However, this is not fair. The problem is that this objection can be raised against every demarcation proposal for the limits of intentionality, no matter how plausible or appropriate it

[81] Notice that the counterfactuality robustness of intentional explanations lies on the variation of proximal stimuli, not on the presence of the distal feature which remains invariable. But there is no actual-sequence explanation of the system's behaviour by appealing to the presence of the distal feature because what ultimately triggers the production of the system's state are the proximal stimuli, not the distal feature (e.g., when there is no light reaching the retina, there will be no tokening of the internal state and so no responsive behaviour).

may be. After all, for every demarcation proposal which specifies whatever property *p* as a minimal condition for intentionality (e.g., causal independence, employment of constancy mechanisms, etc.), it is always possible to object that this proposal does not draw a line between intentional and non-intentional states, but rather a line between *p* states and *non-p* states. Then, the whole debate on the limits of intentionality would become sterile, incapable of yielding any fruit.[82] Here I take a different route. I develop an objection that applies to both the constancy mechanism proposal and the counterfactual robustness argument. It is based on the fact that the distinction between proximal and distal features comes in degree which gives rise to a serious problem for the constancy mechanism proposal in general and the counterfactual robustness argument in particular.

But before developing this objection, a clarification is necessary. I have formulated the constancy mechanism condition as establishing that in order for the relevant state to represent an external condition, it is required that the state still represents the external condition despite the variety of proximal stimuli that reach the system's apparatus. One may wonder if such formulation is compatible with reductionist naturalist theories of mental representation since it appeals to the very notion of representation. In fact, it is very easy to formulate the constancy mechanism condition in non-representational terms: the producer system should employ a constancy mechanism in the production of the state such that there should be a *constancy* between the tokening of the state and the presence of the external condition, despite the variety of proximal stimuli reaching the system's apparatus. That is, the production system should guarantee a *conjunction* between the tokening of the state and the presence of the external condition. That said, this is my fundamental objection to the constancy mechanism proposal.

The counterfactual robustness argument assumes that there is no intentional explanation of behaviour which posits proximal content representations and so that there is no

[82] Marc Artiga has also reached this conclusion but on different grounds, cf. ARTIGA, 2016, p. 422.

representation of proximal features. In other words, it assumes that there is no behaviour prone to proximal content intentional explanations. That is the reason for this conclusion. The fundamental assumption of the argument is that it is a distinctive feature of intentional explanations that they give an account of counterfactual cases. This assumption implicitly assumes that there is no proximal content intentional explanation since it is not possible to give an account of counterfactual scenarios based on the positing of representations with proximal content. Only intentional explanations that posit distal content representations give an account of counterfactual scenarios because here the employment of constancy mechanisms is required in order for the state to represent the same distal feature despite proximal stimuli variations. But if the content of a representation is the proximal feature, what kind of variation could there be among counterfactual situations in order for a proximal content intentional explanation to give an account of them? There is none. An illustration of how distal content explanations give an account of counterfactual situations makes this point clear.

Let's come back to the case of the animal's avoidance behaviour to escape from predators. The animal still represents the same distal feature, the presence of the predator, despite differences among the predator's space positions or in environmental lighting reaching the retina. Such variations constitute the counterfactual scenarios that the distal content intentional explanation is capable of giving an account. It provides explanation of the animal's avoidance behaviour throughout all these counterfactual situations and that is the reason that this intentional explanation is a robust-process explanation, not an actual-sequence one. But what varieties could there be for a proximal content intentional explanation to give an account? None. There is no variation that could constitute any counterfactual scenario which a proximal content intentional explanation could give an account. Hence, proximal content explanations are not robust-process explanations.

Finally, the defender of the counterfactual robustness argument may justify from an

explanatory point of view that there is no genuine proximal content explanation in the following way. The positing of a system's internal state as a proximal content representation in order to explain behaviour makes no explanatory difference since a non-intentional explanation that posits no representation has the same explanatory power than a proximal content explanation. So, only representations of distal features have explanatory power and since only systems with constancy mechanisms can represent distal features, the conclusion is that their employment is a minimal condition for intentionality.

So far, so good. But here a serious problem arises for both constancy mechanism proposal and counterfactual robustness argument – the *minimal distance problem*. Its starting point is that the distinction between proximal and distal stimuli comes in degrees – it is not clear-cut. There are no strict groups of proximal and distal stimuli because there is no non-arbitrary line which strictly divides proximal and distal stimuli. Evidently, certain stimuli are more proximal than others, in the sense that there is a hierarchy of stimuli classified by proximity. But any strict line drawn in order to divide it into two groups – proximal and distal stimuli – will always be blurry. Someone could propose that there would be not two groups, but three groups – distal, intermediary and proximal stimuli. However, in this case, the problem of drawing the limits strikes again, because the distinction between distal and intermediary stimuli and the distinction between intermediary and proximal stimuli are also blurry.

The fact that the distinction between proximal and distal stimuli is not strict raises a problem to both the counterfactual robustness argument and the constancy mechanism proposal: what is the minimal distance from the system's sensory apparatus that a stimulus may be in order for a state to genuinely represent it? Notice that the employment of a constancy mechanism presupposes that there is *some* distance between the representational state and what is being represented. Otherwise there would be no distance based on which a constancy mechanism would keep the state representing the same feature despite huge variations in the

stimuli reaching the system's sensory apparatus. But what is the extent of this minimal distance? The counterfactual robustness argument gives rise to one response, while the constancy mechanism proposal gives rise to another one. I will assess each response by turn.

Maybe the direct response to this problem is that the minimal distance between the stimulus and the token state for the later to represent the former is this one. The stimulus is any stimulus but the most proximal one – the most proximal stimulus cannot be represented. The main argument for this response comes from the constancy mechanism proposal itself. No state can represent the most proximal stimulus because in this case no employment of a constancy mechanism is possible. There would be no distance for the state to represent the same feature despite the stimuli variation reaching the system`s sensory apparatus. Some minimal distance between the tokening of the state and the trigger stimulus is required for the employment of constancy mechanism to be possible in the first place. Furthermore, notice that in the case of the most proximal stimulus which triggers a certain state, it would be explanatory idle to posit that the state represents this stimulus because the non-intentional explanation behaviour would have the same explanatory power. Therefore, the minimal distance between the stimulus and the token state cannot be the most proximal one.

What about the second most proximal stimulus? The counterfactual robustness argument gives rise to the following justification for the second most proximal stimulus constituting the minimal distance, in contrast with the most proximal stimulus. The second most proximal stimulus is the closest stimulus to the sensory apparatus that a given stimulus may be that still gives rise to a robust explanation of behaviour, i.e., an explanation that gives an account of counterfactual situations. The explanation that posits the state as representing the second most stimulus in the stimuli chain gives an account of counterfactual situations. By contrast, an explanation that posits the state as representing the most proximal stimulus cannot give an account of counterfactual situations. Therefore, the latter is not a robust explanation

and thus not an intentional explanation. This is how the counterfactual robustness argument gives rise to a justification for the minimal distance being the second most proximal stimulus, not the most proximal stimulus.

However, this justification is problematic. At the second most proximal stimulus level the group of counterfactual situations is so small and irrelevant that threatens the viability of the claim that there is a genuine robust intentional explanation of behaviour in contrast with non-intentional explanations that lack such robust character. Note that according to the counterfactual robustness argument, it is the amount of counterfactual situations that a given explanation is capable of giving an account which determines whether the explanation is sufficiently robust or not. But if the amount of counterfactual situations is so small, an explanation which gives an account of it is not explanatorily relevant in order to make this explanation genuinely robust.

It could be replied that there is at least one counterfactual situation that an intentional explanation gives an account because it posits that the state represents the second most proximal stimulus, while a non-intentional explanation cannot give an account of this counterfactual situation. However, the fact that an explanation gives an account of just one counterfactual situation is not sufficient to make it robust. The explanation is not sufficiently robust to justify the positing of the state as representing the second most proximal stimulus precisely because this positing is not explanatorily relevant.

The following example illustrates what is problematic with the above justification for the second most proximal stimulus as constituting the minimal distance. Suppose that stimuli chain triggers the tokening of a certain state in a cognitive system and that s^1 is the most proximal stimulus of the chain, s^2 is the second most proximal stimulus, s^3 is the third one and so on. As previously shown, the state cannot represent s^1 because there is no distance between the sensory apparatus and the stimulus. Now consider the hypothesis that the state represents

s^2 because an intentional explanation that posits this state as representing s^2 provides an account of the counterfactual situations constituted of variations of s^1. However, the amount of counterfactual situations constituted by variation of s^1 is so small and irrelevant for the explanation of behaviour that an account of it is not sufficient to make robust an explanation which gives an account of these counterfactual situations. Hence, the positing of the state as representing s^2 is explanatory irrelevant and according to the counterfactual robustness argument the minimal distance is not s^2.

What about the hypothesis that the minimal distance is s^3? The same problem arises again. The amount of counterfactual situations constituted by variation of s^1 and s^2 is still small and explanatorily irrelevant. An account of it is insufficient to make robust an explanation that gives an account of these counterfactual situations by positing the state as representing s^3. Once again, the positing of the state as representing s^3 is explanatory irrelevant.

What is the lesson we should draw from this exercise? The conclusion is that it is not clear what is the first stimulus in the causal chain (s^2, or s^3, or, s^4...) for which the positing of the state as representing this first stimulus gives rise to a robust intentional explanation. That is, an explanation which provides an account of a sufficiently big and relevant amount of counterfactual situations constituted by variations in the intermediate stimuli between the tokening state and the first stimulus in that causal chain. In the absence of some justified criterion, it is indeterminate what is the minimal distance from the sensory apparatus that a stimulus may be for the state to genuinely represent it. This is *the indeterminacy objection* for the response that the counterfactual robustness argument gives rise to the minimal distance problem.

Let's move to the response which the constancy mechanism proposal gives rise for the minimal distance problem. The constancy mechanism proposal maintains that the employment of a constancy mechanism is a minimal condition for a certain state to be representational.

Accordingly, it is a condition for a state to represent a given feature that it represents the same feature despite significant variations of proximal stimuli reaching the system's sensory apparatus. Based on this proposal, it could be argued that since there are variations of s^1 between the sensory apparatus and the second most proximal stimulus s^2, then the room is open for the employment of a constancy mechanism and hence s^2 is the minimal distance. However, this move is problematic. In this case, the only possible variation of proximal stimuli reaching the sensory apparatus is constituted of variations of s^1. Is it enough variation for the state to genuinely represent s^2? Maybe not. It is not clear that there is enough variation for the employment of the constancy mechanism to be possible in the first place – maybe it is only possible when there is *more* variation. The same problem arises for the hypothesis of s^3 as the minimal distance. The only possible variation of proximal stimuli is constituted of variations of s^1 and s^2. Is it now enough variation for the state to genuinely represent s^3? Once again, maybe not. It is also not clear that there is enough variation for the employment of the constancy mechanism. The conclusion is that it is indeterminate what is the minimal distance between the sensory apparatus and the stimulus in order to have enough intermediate stimuli variation. That is, enough intermediate stimuli variation for the employment of the constancy mechanism to be possible in the first place. It is an open question what is the first stimulus in the causal chain for there to be enough variation of intermediate stimuli required for the employment of the constancy mechanism.

The indeterminacy objection threatens both responses to the problem of minimal distance, the response based on the counterfactual robustness argument in particular and the response based on the constancy mechanism proposal in general. Its neutralization requires a non-arbitrary criterion to determine the extent of the minimal distance from the system's sensory apparatus that a stimulus may be in order for the state to genuinely represent it. If the minimal distance is indeterminate, then the constancy mechanism proposal and the

counterfactual robustness argument would collapse. In the absence of a solution to the minimal distance problem, the viability of the constancy mechanism proposal is threatened since it depends on a strict distinction between proximal and distal stimuli. That is, if this distinction is indeterminate, it follows that the constancy mechanism proposal's assumption that there is only representation of distal features is also indeterminate and so the constancy mechanism proposal is in trouble.

At this point, it could be replied that the indeterminacy objection runs the risk of being a slippery slope argument. This risk is illustrated by this analogy. Baldness is a vague notion: it is not fully determinate the point in which someone that is not bald becomes bald. How many strands of hair should a non-bald person lose to become bald? There is no strict turning point, but a continuum between full baldness (no hair) and full non-baldness (full head of hair), somewhere in the middle lies the limits of baldness. Now suppose that a supermodel agency prohibits bald male supermodels. It would be nonsense to claim that this is not a viable condition for supermodel selection since it does not provide a principled way of determining exactly when someone non-bald turns bald. After all, baldness is a useful notion and it works quite well for the selection of supermodels. Analogously, the constancy mechanism proposal cannot be rejected because it does not provide a principled way of determining the extent of the minimal distance from the system that an external condition may be in order for the state to genuinely represent it. Rather, there is a continuum between representationality and non-representationality in which there are clear cases of representational states and clear cases of non-representational states. There is nothing problematic about it just like there is nothing problematic about the continuum between baldness and non-baldness.

In fact, there is nothing problematic with the notion of baldness in its use on our everyday life to distinct bald from non-bald people. Nevertheless, things are different in respect of the philosophical debate on the minimal conditions for intentionality. Here we are looking

for minimal conditions for intentionality to develop a categorisation of representational and non-representational states, not for a continuum between these states. In order to develop such categorisation, the categories of representationality and non-representationality cannot suffer from this level of indeterminacy precisely because it threats the viability of the delivered distinction between representational and non-representational states. Furthermore, notice that the constancy mechanism proposal requires "sufficient variation" in proximal stimuli between the state and the distal external condition for the first to represent the latter (STERELNY, 1995, pp. 261-2). But how huge should this variation be for it to constitute sufficient variation for intentionality? The very specification of this proposal makes it clear that it is not whatever variation that is required for intentionality. That is another reason for the conclusion that indeterminacy of the required amount of stimuli variation is problematic for the viability of this proposal. It is not a solution to this problem to just claim that there is a continuum between representational and non-representational states.

What is the lesson one should draw from this conclusion to the overall debate on the problem of demarcation? I think it is that the appropriate proposal for minimal conditions for intentionality should not rely on a distinction between proximal and distal stimuli unless this proposal has a solid criterion to respond for the minimal distance problem. As result, the objection of indeterminacy would constitute no threat to this proposal. However, it is hard to see how this is a viable way out for the constancy mechanism proposal since it draws the limits of intentionality appealing only to the constancy mechanism condition. Another minimal condition is required to determine the extent of the minimal distance from the system's sensory apparatus that a stimulus may be for the state to genuinely represent it. In the next chapter, I develop a proposal that adopts a variation of the constancy mechanism condition; I show that this proposal is not threatened by the objection of indeterminacy. That is the case because it also adopts another minimal condition that provides a non-arbitrary criterion to determine such

minimal distance – the *success pattern condition*. The result will be the *dual proposal for the minimal conditions for intentionality*.

What about the intuitiveness of the constancy mechanism proposal? Is it compatible with the intuitive principles that are highly entrenched in our intuitive conception of representation? It is clear that the constancy mechanism proposal respects the complexity principle. The employment of a constancy mechanism by a given system guarantees that it has a certain cognitive complexity since the capacity to employ a constancy mechanism makes it more complex than a system deprived of such capacity. The employment of a constancy mechanism makes possible for the system to represent the same external condition despite a variety of proximal stimuli reaching its sensory apparatus. But is the cognitive complexity required for the employment of constancy mechanisms enough for the satisfaction of the complexity principle? That is uncontentious in the case of more complex organisms like humans, chimps and even vervet monkeys. Let's consider simple systems like honeybees to assess if their actual degree of complexity satisfies the complexity principle. The employment of the constancy mechanism guarantees that the organism's behavioural output is not chained to a given stimulus, i.e., a variety of input stimuli may trigger the tokening of the representational state. This fact makes the organism cognitively more complex than organisms like tropistic systems that are chained to a specific stimulus. The employment of constancy mechanisms makes the honeybee complex enough to produce representational states with shape and colour constancy, the behavioural output of foraging is not chained to a specific shape or colour of flowers (the sources of nectar). I take this stimulus-behaviour unchained feature to be complex enough for the systems that share it to satisfy the complexity principle and hence to be genuine candidates for intentional systems.

The constancy mechanism proposal also respects the autonomy principle according to which intentional systems are autonomous, the forces responsible for their behavioural outputs

originate within the systems, not outward. The employment of the constancy mechanism requires autonomy from the system. By producing a state that represents the same external feature despite the variety of input stimuli, the system keeps the tokening of the state constant whenever the relevant external feature is present which triggers the behavioural output. That requires a selection process of picking from all input stimuli just the ones that correlates with the external feature, ruling out the remaining stimuli, which results in the establishment of a constancy relation between the tokening of the state and the presence of the external feature. The selection process that tokens the internal state constitutes an internal force that affects the resulting behaviour. If the selection process tokens the state that triggers the behavioural output, then this process constitutes a force that affects this behaviour. The forces responsible for it cannot be wholly originated outside the system. The conclusion is that the constancy mechanism proposal respects the autonomy principle. The satisfaction of the constancy mechanism condition guarantees the autonomy of the system.

However, the constancy mechanism proposal keeps the door open for the violation of the activity principle. This one requires that the system should have some active role in order for it to be genuinely intentional. But it is possible that even though the tokening of the state is triggered by a variety of proximal stimuli and correlates with the relevant external condition, the system has no relevant active role in the production of the behavioural output. That is the case because a given system that employs the constancy mechanism may have always the same behavioural output as a response to the presence of the external feature. But it is hard to argue that a system that always produces the same behaviour as a response to the presence of the external feature has some relevant active role. The door would still be open for a system which always produces the same behaviour to have some active role if the constancy mechanism condition is conjoined with some other minimal condition that would justify the claim that this system is active even though it always produces the same behaviour. The constancy mechanism

condition alone fails to do that. So, the employment of the constancy mechanism fails to guarantee that the system has some relevant active role. The conclusion is that the constancy mechanism proposal keeps the door open for the violation of the activity principle.

In sum, the employment of the constant mechanisms guarantee that the relevant system respects the complexity and autonomy principles but keeps the door open for the violation of the activity principle. Hence, the constancy mechanism proposal is not entirely compatible with the intuitive conception of representation. In the context of mutual adjustments between intuitive and explanatory constraints on the assessment of minimal conditions for intentionality, it could be argued that putting aside this counter-intuitive aspect of the constancy mechanism proposal is justified because this proposal is explanatorily justified. However, such a move is not allowed for the proponent of the constancy mechanism proposal because as previously showed, this proposal fails to specify the distinctive explanatory power of representations and so is not explanatorily justified. In conclusion, the constancy mechanism proposal should be rejected.

Conclusion

In this chapter, I addressed the problem of demarcation and the minimal conditions for intentionality. I started by defending the substantiality of this problem by arguing that this is not a purely terminological debate on what we call representations, but a solid problem on the nature of intentionality. After that, I have proposed the method of reflexive equilibrium for establishing the appropriate proposal for the limits of intentionality. I argued that the proposal should be constrained by mutual adjustments between explanatory and intuitive requirements in order to reach an optimal equilibrium state in which we are no more inclined to revise the resulting minimal conditions. Finally, I assessed and rejected the causal independence and the constancy mechanism proposals for minimal conditions for intentionality in terms of both

explanatory and intuitive considerations. This concludes the negative stage of my approach to the problem of demarcation. Let us pass on to the positive stage. The goal of the next chapter is to develop and defend my own proposal for minimal conditions for intentionality – the *dual proposal*.

CHAPTER 4. THE MINIMAL CONDITIONS FOR INTENTIONALITY: THE DUAL PROPOSAL

4.1 The success pattern proposal

4.2 The objection of liberality

4.3 The dual proposal: constancy mechanism joins success pattern

4.4 Is the dual proposal intuitive?

The goal of this chapter is to develop the dual proposal for minimal conditions for intentionality, my solution for the problem of demarcation. In the first section, I develop the success pattern proposal. Its starting point is the assumption that intentional explanations have the distinctive power of explaining success, i.e., to explain when certain behaviours are successful or not. Evidently, the success or failure of a behaviour consists in the achievement or not of a given result. Note that success assessments of behaviour by intentional explanations presuppose that the system pursues a given result via the production of this behaviour. This feature of intentional explanations gives rise to the success pattern condition for minimal intentionality: a given system is intentional only if (a) what I call a "success pattern" is present in its behaviour; (b) it uses the representational state as proxy for the presence of the relevant external condition. Nevertheless, in the second section I show that the success pattern condition alone draws the lower border of intentionality too low in light of both explanatory and intuitive considerations. The delivered demarcation is too liberal because it treats systems that are clearly not representational as representational. That is the objection of liberality. The establishment of other minimal condition for intentionality is hence required. In the third section, I propose a variation of the constancy mechanism condition. Together with the success

pattern condition, this constancy mechanism condition constitutes *the dual proposal* for the minimal conditions for intentionality which establishes the genuine limits of intentionality. Finally, I show that even though the dual proposal adopts a variation of the constancy mechanism condition, it is not threatened by the indeterminacy objection. This objection is problematic for the constancy mechanism proposal, but not for the dual proposal. The reason is that the latter nonarbitrarily determines the minimal distance from the sensory apparatus that the external condition may be for the system to genuinely represent it.

In the last section, I show that the dual proposal is compatible with the intuitive conception of representation to an acceptable extent because it respects three of the most entrenched intuitive principles on representationality – the activity, autonomy and complex principles. Beyond this extent, however, the limits of intentionality become blurry. It turns out to be a pragmatic choice to accept or reject the intentionality of some states. The reason is that it is not possible to establish a fully precise demarcation without violating the explanatory requirement. The result is the optimal state in the reflective equilibrium. The mutual adjustments between the distinctive explanatory role of representational states and our intuitive conception of representation achieves the optimal equilibrium. There is no inclination to revise the resulting proposal anymore, i.e., the dual proposal.

4.1 The success pattern proposal

Representational states by themselves cannot make any difference for the organism or the external environment. An isolated representation is idle unless it triggers other state(s) that will trigger behaviour. It is the production of behaviour, either directly by triggering a behavioural state or indirectly by triggering other internal state(s), that may or may not achieve

any supposed pursued goal of a representational state.[83] Hence, the achievement of the organism's pursued external result is only reachable through the behaviour triggered by the representation. The success pattern proposal maintains that the positing of a representational state gives rise to an intentional explanation with the distinctive explanatory power of *explaining success*. That is, the intentional explanation specifies the external condition in which the triggered behavioural output succeeds in achieving the pursued result. By contrast, the non-intentional explanation does not appeal to any pursued result and hence cannot explain successful behaviour. That is the fundamental distinction between intentional and non-intentional explanations according to the success pattern proposal. But what is the precise nature of the distinctive explanatory power of the intentional explanation in explaining successful behaviour? How does the fact that the representational state pursues a certain external result make any difference to the explanation of behaviour?

Let's start this investigation by taking a look at the extent of the explanatory power of non-intentional explanations. Suppose that a given organism has a behaviour pattern *B* in response to a certain internal state *S* that is triggered by an external condition *C*. What could be the explanation for such behaviour? The non-intentional explanation consists in a specification of the causal chain of the tokening of *C* that triggers *S* which results in *B*. It is potentially capable of explaining every causal transaction that occurs between the tokening of *C* and the production of *B*. In the ultimate and ideal case, the non-intentional explanation consists in a complete explanation of the causal chain that specifies every single atomic and molecular movement that occurs between the tokening of *C* and the production of *B*. It is potentially capable of explaining every single bodily movement of the organism by providing an explanation of every causal transaction that originated it. As is shown by the sciences of

[83] Is it really possible for a representational state to directly trigger behaviour? That is, without the intermediation of other internal state(s) that will produce behaviour? That possibility makes no difference for the ongoing argument and so I will just leave that door open.

mind after the cognitive revolution, a full causal explanation of behaviour is made possible by providing a wiring diagram that shows how environmental inputs affect internal states of the organism that, in conjunction with other internal states, originate the behavioural output. Finally, it is typically assumed by cognitive sciences that representational states interact casually in virtue of their non-semantic properties (i.e., syntactic properties) to originate behavioural outputs. Nevertheless, these causal transactions are faithful to the content of the representational states. This is well illustrated by Dretske's analogy that a glass may shatter if you scream "shatter", but your screaming's meaning is causally irrelevant to the shattering (DRETSKE, 1988).[84] In light of this, a serious problem arises for the justification of intentional explanations: given that the non-intentional explanation has the power of providing a full causal explanation of behaviour, what would be left to be explained by an intentional explanation? What is missed by the non-intentional explanation that could justify the positing of *S* as a representational state?

The starting point is to notice that intentional explanations are external explanations, while non-intentional explanations are internal explanations. The causal explanation of behaviour establishes no relation or connection between the organism and the external environment; it is true of the organism irrespective of its external environment. If everything is changed in the external environment with the exception of the proximal stimulus that triggered the causal chain that resulted in the bodily movement, the causal explanation would still be true of the organism. By contrast, intentional explanations are external explanations. The intentional explanation establishes a new set of relations between the organism and the external environment where it is embedded by establishing relations between representational

[84] There is a lively debate on whether representational content has causal relevance or efficiency. Some philosophers maintain that content is not causally inert and hence that representational states causally interact in virtue of their content, while others disagree (cf. DRETSKE, 1988; BLOCK, 1989; RESCORLA, 2015). However, it is not my focus to enter into this debate. Here I am just arguing that there is a complete non-intentional explanation of behaviour, that is, a full causal and non-semantic explanation of behaviour.

states and external features of the environment (SHEA, 2013, p. 498). The non-intentional explanation only specifies the relation between the organism and a certain proximal feature of the environment, namely, the proximal stimulus that triggered the relevant causal chain. Once it has specified this proximal stimulus, the non-intentional explanation has nothing more to say about the external environment. Hence, the establishment of relations between the representational state and some environmental external features is distinctive of intentional explanations.

But how does the intentional explanation establish a relation between the representational state and the external feature of the environment? It comes in two steps. First, the intentional explanation specifies the result or end that the organism pursues with the tokening of the representation and the resulting behavioural output. The pursued result is directed towards a given external feature in the environment – e.g., escaping from something, eating something, approaching something, etc. This is the reason that it is an external result, not an internal one. Second, the intentional explanation specifies the success conditions for the behavioural output to achieve this result. The establishment of the relation between the organism and the external feature comes in two steps because the establishment of success conditions presupposes the specification of the pursued result; otherwise, it would not be possible to establish whether the behaviour succeeds in achieving it or not.

The distinctive explanatory power of intentional explanations consists in the explanation of successful behaviour, i.e., the prediction of the success or failure of the behavioural output in achieving the organism's persuaded result. The establishment of success assessments of behaviour is what gives rise to intentional explanations of behaviour. The intentional and non-intentional explanations establish different descriptions of behaviour at different levels, namely, the success pattern description and the non-intentional description. No matter how complete the non-intentional specification of the causal chain that leads from the

relevant proximal stimulus to the behavioural output, it is not capable of explaining the success or failure of the organism in achieving the pursued result. That is, the non-intentional explanation is not capable of explaining successful or unsuccessful behaviour. We can illustrate it with the predator example.

Suppose that a given organism pursues the result of avoiding predators and that the tokening of a certain internal state is what triggers a certain behaviour. As it is clear, the pursued result is towards the presence of predators – after all, the organism is pursuing their avoidance. Now suppose that in a given situation the tokening of this internal state triggers the behaviour which may lead to the avoidance or not of a predator. The non-intentional explanation of this behaviour is the specification of the causal chain that starts with the stimulus input that leads to the token behaviour and ends with the behavioural output. But what is the point of positing that this internal state represents the predator? Because it gives rise to an explanation of the success or not of the organism in achieving the relevant result, i.e., escaping from the predator. The true representation of the predator leads to the avoidance behaviour which leads to the success of escaping from the predator; the false representation leads to the failure of avoiding the predator simply because there is no predator around. When the state is true, it contributes for the organism to avoid predators; when the state is false, there is no contribution; on the contrary, it leads to loss of energy. The intentional explanation is required precisely for the explanation of the success or not of escaping from the predator. Two factors guarantee the success of the organism in escaping from the predator. First, that the internal state truly represents the presence of the predator – the external condition of the presence of the predator obtains; second, that the behavioural output is the avoidance behaviour. The truth of the representation guarantees success provided that the behavioural output is the appropriate one. After all, in the case that the organism does not exhibit avoidance behaviour, it would make no difference whether it truly represents the presence of the predator or not.

The explanatory power of intentional explanations lies in the capacity to explain success, that is, the achievement of the organism's pursued result. There is a characteristic pattern assumed by intentional explanations, the success pattern – a certain appropriate behaviour *B* in pursuing result *R* is prompted by representational state *S* that represents external condition *C*; when *S* is true (i.e., when *C* obtains), the performance of *B* leads to the achievement of *R*. The truth of the representation explains the success of the appropriate behaviour to achieve the pursued result and the falsehood explains the failure of achieving this result. The positing of the representational state by the intentional explanation provides it with two explanatory powers that are absent in non-intentional explanations – *generality* and *prediction*. That is precisely what justifies the positing of representational state and the appeal to the intentional explanation.

This is the success pattern proposal of the explanatory power of intentional explanations. Its core thesis is that the distinctive explanatory power of intentional explanations is the capacity to explain success. It was originally proposed by success semantics (RAMSEY, 1927; WHYTE, 1990) and it is also assumed by the main proponents of teleosemantics (MILLIKAN, 1984, 2004; PAPINEAU, 1993, 2016). However, the success pattern proposal was not originally formulated as a solution for the problem of demarcation, my job here is *to adapt it* as a proposal for the minimal conditions for intentionality in the context of this problem. The success pattern proposal identifies success conditions with truth conditions – the representations is true if and only if it contributes to the achievement of the pursued external result, and false otherwise. The truth condition is the condition that should obtain for the prompted appropriate behaviour to be successful, i.e., to achieve the result.

Here it is necessary to make an observation on the conception of behaviour assumed above. One may object that even assuming that the representational state has a pursued result (and so that there is a success assessment of the behavioural output), there is nothing left to be

explained by the intentional explanation once there is a full causal explanation of behaviour. Rather, intentional explanation doesn't explain behaviour, but success. I think that what is at issue here is just a terminological matter of what is being called "behaviour". This objection assumes that behaviour is just bodily movement and hence there is nothing left to be explained by an intentional explanation once there is a full explanation of bodily movement. After all, the full causal explanation specifies the causal transactions that result in the bodily movement. However, I am assuming a broader notion of behaviour according to which a complete explanation of behaviour is not restricted to a full causal explanation of behaviour. Rather, a complete explanation of behaviour also explains when such bodily movement is successful or not. That is, it also contains an explanation of successful behaviour. This is the sense in which intentional explanations are explanations of behaviour. Evidently, one can choose to call "behaviour" only the bodily movement and hence to claim that the subject matter of intentional explanations is not behaviour, but success. Nevertheless, this is only a terminological matter of what is being called "behaviour", which does not touch in any way the viability of the success pattern proposal of intentional explanations.

So far, so good. However, an ontological question remains about the nature of the success pattern assumed by this account of intentional explanations – what is the ontological nature of the success pattern? On the one hand, there is the anti-realist response according to which the success pattern is not real; it does not really exist in the behaviour. It is just a projection on the behaviour made by the observer – once there is no one observing the behaviour, there is no success pattern. By contrast, the realist response claims that the success pattern exists and is present in the behaviour independently of whether there is an observer around or not. There is no projection being made, the success pattern is not on the eyes of the observer, rather the observer just discoveries a previously existing pattern. In what follows, I will argue that success patterns are real in the sense of real patterns introduced by Daniel

Dennett (DENNETT, 1991).

Success patterns are patterns of interaction between the behavioural output of the organism and the input stimuli from the external environment. To say that a given success pattern exists in an organism is to say that the behavioural output of the organism exhibits such a success pattern. A success pattern is real because the behavioural output has such a pattern irrespective of whether there is anyone who recognises that it is there. Rather, there would be a success pattern there even if an observer had never existed who recognises that it exists. For instance, consider the catching behaviour of an organism towards food in different situations where food is located in different places and nevertheless the organism catches it. The intentional explanation for it is that the organism represented the presence of food and so catches it. The success pattern is on the catching behaviour throughout these situations irrespective of whether there is an observer to verify the presence of this pattern on these occasions. Evidently, for an observer to recognize the existence of a success pattern in the behaviour of a given organism it is required for the observer to posit an internal representational state on this organism. But the intentional explanation does not consist in a projection made by the observer of a pattern that would not exist in case of the absence of the projection; rather, the success pattern is present in the behaviour irrespective of any projection. There is a fact of the matter as to whether a given organism's behavioural output has a success pattern, irrespective of whether there is someone to recognize that the behaviour exhibits such a pattern. Intentional explanations pick up real success patterns of the way the organism interacts with the surrounding environment; they are there independent of any projection or observation.

The success pattern proposal was originally proposed by success semantics to give an account of intentional explanations of behaviour given rise to by belief and desire states, but it is generalizable to every intentional explanation – including intentional explanations given rise to by representations with minimal intentionality. However, the success pattern proposal comes

with a problem: how do intentional explanations determine the external result that the organism pursues? There is an instance of a success pattern only if the relevant organism pursues a given result, but how does the intentional explanation determine which the result is? For instance, in the predator example, what is the criterion for the specification that the organism's pursued result is the avoidance of a predator but not a different result?

Success semantics originally appealed to the success pattern proposal in order to determine the truth-conditions of beliefs. A belief's truth-condition is the condition that guarantees the fulfilment of any desire by the behaviour which that belief and desire would combine to cause.[85] But such specification of the belief's truth-conditions presupposes the specification of the desire's truth-conditions – it determines the content of beliefs in terms of the content of desires. Thus, it provides no solution for the problem of how to determine the organism's pursued result (in the case of belief-desire systems, the satisfaction of the desire), it just presupposes it. Furthermore, such specification of the belief's content is not acceptable in light of the naturalistic standard since it presupposes the semantic notion of the desire's content. The case of success semantics illustrates quite well the problem of appealing to the success pattern in order to give an account of the explanatory role of intentional explanations: how to specify the pursued result in a naturalistically acceptable way?

It is here that teleosemantics comes into play. Teleosemantics claims that the pursued result of a representational state is constituted by its biological goal or purpose. The success of achieving the pursued result is just biological success. The organism tries to fulfil a certain biological function via the tokening of the representational state and the triggered behavioural output. It is such biological function that determines the organism's pursued result and, derivatively, the representation's pursued result. The representational state is teleological, it

[85] Cf. WHYTE, 1990, p. 150. Frank Ramsey, the founding father of success semantics, determined the truth-conditions of a belief in the following way: "any actions [behaviours] for whose utility [success] p is a necessary and sufficient condition might be called a belief that p, and so would be true if p, i.e. if they are useful [successful]." (RAMSEY, 1927, p. 5).

has a biological function that determines its success conditions and hence the success conditions of the triggered behavioural output in achieving this pursued result, that is, its biological goal. There are several theories of biological function that specify how a given representational state has a biological function (aetiological theory, systemic theories, etc.)[86], but it is not my focus here to assess them. It is sufficient to note that there are rival theories of biological function. Here I will just assume that one of them is right. Hence, a representational state has a biological function that determines its pursued result and so the success conditions for the resulting behavioural output to achieve its pursued result. In the predator example, the goal of the representational state is to avoid predators because its biological function is to avoid predators. Teleosemantics assumes the core thesis of success semantics of identifying truth conditions with success conditions and takes the further step of specifying the success conditions of representational states in terms of their biological function (MILLIKAN, 1984, 2004; PAPINEAU, 1993, 2016.). In short, success conditions are biological success conditions and the external result pursued by the organism via tokening the representation is its biological goal.[87]

But it should be highlighted that the success pattern proposal is independent of the teleological specification of the pursued result. In fact, it is compatible with any naturalist specification of it, teleological or not. My preference for a teleological specification should not be confused with an endorsement of the thesis that the success pattern is committed to a teleological specification of the pursued result. Note that the fact that the organism's pursued result is specified in terms of biological function (or in any other way) does not play any explanatory role in the success pattern proposal. This proposal just assumes that there is a result

[86] For the aetiological theory of function, cf. WRIGHT, 1973, MILLIKAN, 1989a; NEANDER, 1991. For the systemic theory, cf. CUMMINS, 1975.

[87] Here the following problem arises for the viability of teleosemantics: why should the pursued result be specified in terms of biological function? That is, why should the biological function be the one that specifies the pursued result, not something else? However, it is not my aim to assess this problem here. In what follows, I will just assume that it is the biological function of the representation that determines the pursued result.

pursued by the organism via the tokening of the representation; what determines this result is a further and independent matter. In what follows, unless otherwise specified, I will remain neutral on the question of the appropriate specification of the pursued result.

The success pattern proposal maintains that it is a minimal condition for a given system to be intentional that a success pattern is present in its behaviour. That is, if there is no success pattern in the behavioural output, it follows that the system is not genuinely representational. When the success pattern condition is combined with the teleosemantics identification of the representation's pursued result with its biological function, it becomes the teleological pattern condition. Hence, the teleological pattern condition is just a variation or a subtype of the success pattern condition. Evidently, the conclusion that a system is in fact intentional depends on the verification that its behaviour shows this success pattern, but this condition is not merely stating that the system is intentional only if you can recognise that its behaviour has a success pattern. Rather, it is stating that it is a condition for minimal intentionality that the success pattern is present in the behaviour; how this pattern is verifiable is a distinct and further matter. In fact, different systems may require different methods of verification that vary from the nature of the systems. For instance, the method of verification of the success pattern as present in the behaviour of a person is different from the method of verification of it as present in the behaviour of an animal. That is why cognitive science and neuroethology have different but continuous subject matters and methodologies (the latter studies animals like toads and bees while the former studies humans).

Finally, the success pattern condition does not establish that in order for a given internal state to be intentional, the presence of the success pattern in the behavioural output is enough. It also requires that the organism should use the internal state as a proxy for the presence of the relevant external condition to achieve the pursued result. That is, the internal state guides the production of the behavioural output in the achievement of the organism's pursued result. The

organism has a pursued result that is only achievable when a given external condition obtains. Thus, it uses the internal state that stands for the external condition as a guide for achieving this result via the production of the behaviour. In the predator example, the relevant internal state is a representation of the predator because it satisfies both requirements. First, the behavioural output has a success pattern (the success condition is the avoidance of predators). Second, the organism uses the internal state as a proxy for the whereabouts of the predator to escape the predator (e.g., move in the opposite direction). That is, the internal state guides the organism in the production of the behavioural output in so far as it informs the organism about the location of the predator. In sum, the success pattern condition establishes that it is a minimal condition for intentionality that the behavioural output has a success pattern and that the organism uses the relevant internal state as a proxy for the relevant external condition in the production of the behavioural output in order to achieve the pursued result. The success pattern condition is satisfied if and only if both requirements are satisfied.[88]

4.2 The objection of liberality

The success pattern proposal is a clear and straightforward proposal of the distinctive explanatory power of intentional explanations – the explanation and prediction of successful behaviours, i.e., the achievement of the organism's pursued result by tokening the representational state. Whenever such a condition is not satisfied, the positing of the representational state plays no explanatory role and hence must be discarded. It is clear that the success pattern condition rules out several candidates for representational states. For instance, consider a variation of Dennett's lectern example (DENNETT, 1987, p. 23). There is a stone

[88] Ruth Millikan (1984) is famous for introducing the distinction between the producer system that produces the internal state from the consumer system that consumes the internal state. The part of the organism that uses the state as a guide for the production of the behavioural output in order to achieve the pursued result is the consumer system. However, I will not commit myself here to a clear-cut distinction between these systems; it is enough to claim that the organism uses the internal state as a guide for the production of the behaviour.

in a given place that stays there unless something moves it to another place (e.g., a person or a strong wind). How to explain such a stationary position? The intentional explanation claims that the stone has an internal representational state that represents something in the surrounding environment that keeps this stationary position. Why would it be absurd to appeal to this intentional explanation to explain the stone's stationary position? According to the success pattern proposal, the stone clearly does not have a behavioural output with success conditions. It pursues no result and so its supposed stationary behaviour cannot be assessed in terms of the success or failure of achieving any result. In light of the teleological pattern condition, this conclusion is even more persuasive, since it is clear that the stone has no biological function that could constitute a pursued result. The non-intentional explanation of the causal transactions responsible to keep the stone stationary explains everything that has to be explained about the stone's stationary position. Following the same line of reasoning, the success pattern condition rules out several other absurd candidates for intentional systems.

The success pattern condition also rules out the intentionality of several other systems for another reason. If the internal state is not used by the system as a proxy for some external feature in the production of the behavioural output, then it is not intentional even if the system pursues an external result via the production of this behaviour and hence has success conditions. Notice that the success pattern condition does not claim that the mere presence of success conditions in the behavioural output is sufficient for intentionality. It also requires that the system uses the state as a guide in the production of the behaviour to achieve its pursued result. After all, it is evident that the presence of success conditions on the behavioural output is not sufficient for minimal intentionality. It would include several states that are clearly not representations as representational states. The problem is that success conditions are widespread – artefacts like thermostats have success conditions (namely, to keep the temperature at a given set point), every biological system has success conditions in so far as it

has one or more biological functions, every goal-directed system has success conditions, etc. Thus, the requirement of the presence of success conditions is not sufficient for minimal intentionality.[89]

Several systems are not intentional despite the presence of success conditions on the behavioural output precisely because there is no use of the relevant state as a proxy for the presence of the relevant external condition to achieve the pursued result. For instance, the eagle's wing has success conditions constituted by the biological function of soaring. But the eagle does not use the wing state as a proxy for the presence of some external feature to achieve some biological goal. In fact, there is no use of the wing state as a proxy of any feature at all. The fact that the wing is assessable in terms of the success of achieving soaring is not sufficient for it to be intentional.

So far, so good. The success pattern condition succeeds in ruling out several internal states that are clearly not representational, either because the behavioural output has no success conditions or because the system does not use the relevant state as a proxy for the presence of some external condition to achieve the pursued result. Nevertheless, a serious threat comes with the assumption of the success pattern condition – the objection of liberality.

The objection is that the success pattern condition draws the lower border of intentionality too low. It seems that there are several systems that satisfies the success pattern condition but nevertheless are clearly not intentional. Consider thermostats: devices which sense the environmental temperature in order to keep it at a set point. Thus, whenever temperature is below this set point, the thermostat triggers another device that increases the system's temperature, while whenever it is above the set point, the thermostat triggers the device to decrease the temperature. Why not give an intentional explanation of the thermostat's

[89] Even worse, it is subject to the objection of pansemanticism. Maybe there are success conditions everywhere (why not?) which would lead to the absurd consequence that every state in the universe is intentional.

behaviour and conclude that the thermostat represents the environmental temperature? After all, the thermostat pursues the maintenance of temperature at the set point and it succeeds whenever it keeps the temperature at this set point. Evidently, the thermostat and other artefact cases are easily ruled out as genuinely representational states by appealing to the original intentionality condition – the thermostat has no original intentionality. The designer that constructed the thermostat is the one that pursued the maintenance of the temperature in the relevant set point. Hence, the thermostat would have only derived intentionality and a derived pursued result.[90] But note that what rules out artefacts as genuine representations is the original intentionality condition, not the success pattern condition.

Success patterns are wide spread. Even restricting the success pattern condition to systems that have original pursued results (and hence ruling out systems with derived pursued results), the resulting demarcation proposal draws the limits of intentionality too low. Success patterns are recognizable not only in several artefacts but also in very simple biological organisms – even very implausible candidates for representational systems satisfy the success pattern condition. They not only have behavioural outputs with success pattern, they have behavioural outputs with success patterns and use the relevant internal states as proxies for the presence of the external conditions to achieve the pursued biological goal. This is quite clear in the following example.

When the osmolarity of human blood (roughly, the number of particles in a given quantity of blood plasma) reaches above a certain level, the antidiuretic hormone (also called "vasopressin") is produced by the hypothalamus. It then acts on the kidneys, where it triggers an increase in the quantity of water that is reabsorbed during urine formation, resulting in the lowering of the osmolality of the blood. Once again, the intentional talk arises here because the

[90] Here I assume that thermostats have derived intentionality since they are artefacts. It could be objected that that there are real cases of natural thermostats and hence that these ones would have original intentionality (provided that they are intentional). However, the appeal to thermostats is merely illustrative. Another example of an artefact with derived intentionality would also be appropriate to illustrate the objection of liberality.

success pattern is clearly present on the antidiuretic hormone. Why not claim that the antidiuretic hormone produced by the hypothalamus represents high plasma osmolarity? The hypothalamus is an intentional system because it has a clear pursued result (to keep the osmolarity in a certain level), it produces the antidiuretic hormone as a signal for high osmolarity, and the kidney uses the antidiuretic hormone as a proxy for the high osmolarity to lower the osmolarity of the blood. This "behaviour" is successful whenever it achieves this result. Even worse, in this case the appeal to the original intentionality requirement would not be able to rule out the hypothalamus as a representational system. After all, there would be no derived intentionality here, but an unequivocal original intentionality – the hypothalamus has the biological function of producing antidiuretic hormone to lower the level of osmolarity. The hypothalamus is a system that satisfies not only the success pattern condition, but also the teleosemantic version of the success pattern condition and the original intentionality requirement. The same line of reasoning entails that tropistic systems are also intentional.

Considering the anaerobic bacteria example, Tyler Burge claims that the positing of the magnetosomes state as representational is not explanatorily justified. It is clear that there is a success pattern present on the bacterium's behaviour since (a) it has the biological goal of being attracted to the prevailing magnetic field because it correlates with the direction of oxygenated-free water that is lethal for it; and (b) the bacterium uses the magnetesomes state that triggers this behaviour as a proxy for the direction of the magnetic field. But the bacterium's behaviour is fully explained by the non-intentional explanation that specifies the relevant causal chain between the magnetic field and the bacterium's movement towards the magnetic field, and the positing that the magnetesomes state has the biological goal of avoiding oxygenated waters. So, there is nothing left to be explained by the intentional explanation – the positing of the magnetesomes state as a representation is explanatorily idle. In Burge's words, "everything in the example can be explained using the notion of biological function (with respect to oxygen

poverty), normal environmental conditions, and sensory discrimination (with respect to magnetic forces). Adding an odd use of the term 'representation' contributes nothing to explanation." (BURGE, 2010, p. 300).[91]

What about the intuitiveness of the success pattern proposal? Is it compatible with the intuitive conception of representation or does it violate our highly entrenched principles? I think that the lower border of intentionality drawn by the success pattern proposal is intuitively too low. This conclusion is based on both the fact that it violates highly entrenched principles of our intuitive conception of representation and it also treats several states as representations that are not intuitively representational. First, it violates the complexity principle. According to this principle, a certain cognitive complexity is required in order for a given system to be intentional. It considers microorganisms like bacteria and even unicellular organisms like paramecia as full-blooded representational states. However, whatever is the minimum cognitive complexity required by the complexity principle, microorganisms like bacteria and unicellular organisms like paramecia do not satisfy it. They are cognitively too simple.

Second, the success pattern proposal violates the activity principle. This one establishes that an autonomous system that is deprived of any active role – i.e., a wholly passive system – is not intuitively intentional. What happens is that the supposed representational state is automatically triggered by the presence of the relevant external feature that finally triggers the behavioural output. The success pattern proposal treats the magnetosome state of the anaerobic bacterium as representational, but the bacterium is fully passive in the whole process. The orientation of the bacterium in a given direction (bottom of the ocean) is a direct result of the force exerted by the prevailing magnetic field. The bacterium plays such a passive role that, even dead, it aligns itself with the magnetic field.[92] But if the bacterium is dead, it is not

[91] Kim Sterelny makes similar claims about the explanatory idleness of positing tropistic systems and other very simple biological organisms as intentional systems, cf. STERELNY, 1995.
[92] Cf. BLAKEMORE, 1975, p. 379; SCHULTE, 2015, p. 125.

representing anything. In light of the violation of the complexity and activity intuitive principles, the conclusion is that the lower border of intentionality drawn by the success pattern proposal is not intuitive. This proposal is not justified from either the explanatory or the intuitive perspectives.

What is the lesson to be drawn from these intuitive and explanatory objections against the success pattern proposal? I think that it is that the satisfaction of the success pattern condition is not enough to give rise to intentional explanations. That is, the fact that there is a success pattern present in the organism's behaviour and that the organism uses the relevant internal state as a proxy for the presence of some external condition in order to produce a behaviour to achieve the pursued result is not sufficient for the system to be intentional. If every system that satisfies the success pattern condition gives rise to intentional explanations, it follows that there is an intentional explanation for the osmolarity, the paramecium, the anaerobic bacteria and other cases. Notice that these are different cases from the aforementioned eagle's wing state example in which there is no use of the state by the eagle as a proxy for any external feature. Rather, in these cases there is an effective use of the relevant state as a proxy for the external feature to achieve the biological goal. The positing of the representational state is not explanatorily justified in these cases because the behavioural output is fully explainable by the non-intentional explanation that specifies the causal chain from the relevant stimulus to the behavioural output and the result pursued by the organism. On the other hand, the success pattern proposal violates two of the most intuitive principles entrenched in our intuitive conception of representation: the complexity and the activity principles. The conclusion is that the success pattern proposal is too liberal; it draws the lower border of intentionality too low. Thus, some further minimal condition for intentionality is required to make it more restrictive. That is precisely what I will do in the next section. I will propose the adoption of a second minimal condition that together with the success pattern condition will

constitute my proposal for minimal conditions for intentionality – *the dual proposal*.

4.3 The dual proposal: constancy mechanism joins success pattern

In what follows, I will propose that intentional patterns are a subset of success patterns. Every intentional pattern is a success pattern, but not the other way around. As a result, the behaviours of systems like the hypothalamus and the paramecium have success patterns but not intentional patterns, i.e., they have non-intentional success patterns. The explanatory purchase of representational states is not justified in cases of systems with non-intentional success pattern behaviour, only in the case of systems with intentional success pattern behaviour. Evidently, in order for this distinction to be tenable it is required to establish a criterion based on which the distinction can be made.

Success patterns are real. They are patterns of interaction between the organism's behavioural outputs and a certain external condition. They are present in the behaviour irrespective of whether there is an observer verifying its existence; they are not mere projections of the observer. They are present as long as the organism's behavioural output pursues the achievement of a given result, giving rise to success assessments of the behavioural output in achieving this result. The distinction between intentional and non-intentional success patterns should thus be drawn based on the distinctive explanatory power of intentional explanations. But how to do that?

Success patterns give rise to certain kinds of explanations, namely – explanations of the success or failure of the behavioural output in achieving the pursued result. Since we are looking for a distinction between intentional and non-intentional success patterns, we should look for the distinction in explanatory powers due to intentional and non-intentional success patterns. The right approach to establish this distinction is to take a deeper look at the explanatory powers of the explanations given rise by success patterns in general, and then try

to select among these explanatory powers the ones that are distinctive of intentional explanations.[93]

I have previously defended that the presence of a success pattern endows the explanation of the behavioural output with generalization and predictive powers. However, this is not enough to justify the positing of the representational state. In what follows, I defend that the positing of representational states requires further explanatory powers, that is, explanatory powers that are not given rise to by the presence of the success pattern. I claim that the employment of the constancy mechanism gives rise to such further explanatory requirements because it allows the recognition of the presence of the same pattern despite the variety of input stimuli. A success pattern on the behavioural output of a given organism is intentional only if the organism employs a constancy mechanism on the production of the representational state. The presence of the success pattern guarantees the rising of certain explanatory powers that originate in the behavioural output triggered by the representation, while the employment of the constancy mechanism guarantees the rising of other explanatory powers that originate in that which triggers the tokening of the representation. The fundamental distinction between these explanatory powers is that the first arises from what is caused by the tokening of the representation (output side), while the second arises from what causes the tokening of the representation (input side). They are distinct but complementary explanatory powers. The presence of a success pattern gives rise to an output explanatory power, but it does not guarantee that this success pattern is intentional. The employment of the constancy mechanism gives rise to an output explanatory power that guarantees the intentionality of the relevant

[93] What about a pragmatic approach? The distinctive explanatory power of intentional patterns is the result of a mere pragmatic choice of the scientist that has the preference of explaining certain success pattern behaviours but not others in an intentional way. So, the distinction between intentional and non-intentional success patterns is a pragmatic projection of the scientist in light of their own theoretical interests (simplicity, utility, etc.). However, this proposal goes in the opposite direction to the intentional realism that I have defended. If the distinction between intentional and non-intentional success patterns is pragmatic, then the explanatory power of intentional explanations derives from a pragmatic choice of some success patterns as intentional.

success pattern. This is a variation of the constancy mechanism condition that I have assessed in the third chapter. But now it is applied for the recognition of the same success pattern despite the multiplicity of input stimuli that causes the tokening of the representation.

The satisfaction of the success pattern condition guarantees that there is a success pattern in the behavioural output; the satisfaction of the constancy mechanism condition guarantees that this is an intentional success pattern. In sum, the explanatory powers that are distinctive of intentional explanations and that justify the positing of the representational state are constituted by the explanatory power given rise to by the presence of the success pattern (output explanatory power) and by the explanatory power given rise to by the employment of the constancy mechanism (input explanatory power). In case of the dissatisfaction of either the success pattern or the constancy mechanism conditions, the system is not intentional. Both are minimal conditions for intentionality from an explanatory point of view.

Let's start this investigation considering tropistic systems like paramecia. Tropism is the movement of the system (or parts of it) in given direction, in response to a given external stimulus. Now consider that the presence of a certain stimulus in the surrounding environment makes a given tropistic system to move in the opposite direction. The system moves in the opposite direction whenever such a stimulus is present. There are three different explanations for this avoidance behaviour. First, the causal explanation consists in the specification of the causal chain that starts with the presence of the stimulus and ends with the movement of the system in the opposite direction. So, the system has this avoidance behaviour because of this causal pattern. Second, there is the non-intentional success explanation that specifies the causal chain and posits a goal-directed intermediate state between the stimulus and the avoidance behaviour that produces this behaviour to achieve the pursued result. It consists in the specification of this non-intentional success pattern. So, the explanation is that the system has this avoidance behaviour because it tries to achieve its pursued result – it is through this

behaviour that it tries to achieve this result. The difference is that the causal explanation is just the specification of the causal chain, while the non-intentional success explanation posits a goal-directed intermediate state in the causal chain (i.e., the goal-directed state) to explain the avoidance behaviour. Finally, there is the intentional success explanation, i.e., the intentional explanation. It also posits intermediate goal-directed state between the stimulus and the avoidance behaviour, but in contrast with the non-intentional success explanation it posits an intermediate state that is not only a goal-directed state, but something more – a representational state. The system moved in the opposite direction of the stimulus because it represented its presence and this representational state triggered the avoidance behaviour. So, is the positing of this representational state explanatorily justified? If not, why does it have no explanatory purchase? Why does the intentional explanation have no distinctive explanatory role for tropistic systems?

Suppose that there is a success pattern in the behavioural output of this tropistic system constituted by the biological goal achieved by avoiding the relevant stimulus. So, the distinction between the intentional and the non-intentional success explanation is that the first assumes that this is an intentional success pattern (i.e., the intermediate state is a representational state) and that the second assumes that it is a non-intentional success pattern (i.e., the intermediate state is just a goal-directed state). But from an explanatory point of view, it makes no sense to posit that this success pattern is intentional because the intentional success explanation that it gives rise to has the same explanatory power as the non-intentional success explanation. There would be nothing left to be explained by the intentional success explanation that the non-intentional one has not explained. Given that both explanations have the same explanatory power, it follows that the positing of the representational state should be ruled out.

Now contrast this success pattern in the tropistic system with genuine intentional success patterns. What would a non-intentional success explanation miss? In the case of

intentional success patterns, the representational state is triggered by several stimuli (s^1, s^2, s^3…) that triggers behavioural outputs to pursue the same result. In all of these cases, the intentional explanation bridges across a variety of different input stimuli (s^1, s^2, s^3...). The positing of the representational state allows the intentional explanation to identify the presence of the same success pattern despite the variety of input stimuli. By contrast, the non-intentional success explanation misses the identification of the same success pattern that is present throughout all of these cases because they involve different causal chains that start with different input stimuli. For every different causal chain, the non-intentional explanation provides a distinct and independent explanation of behaviour. It misses the same success pattern that is common between them. It is true that you can sum up all distinct causal chains into one single causal explanation by simply conjoining them. But such movement would not allow the identification of the same success pattern, the mere conjunction of distinct causal explanations does not make any connection between them. The identification of the same success pattern is only possible via the positing of the intermediate state as a representation. Such positing by the intentional explanation allows for the recognition of the same success pattern that is present in all of these distinct causal patterns. That is, such recognition is only possible because the organism employs a constancy mechanism in the production of the representational state that allows it to represent the same external feature despite the variety of input stimuli reaching the sensory apparatus. This is a variation of the constancy mechanism condition because the employment of the constancy mechanism is now applied for the recognition of the same success pattern despite the multiplicity of input stimuli that cause the tokening of the representation.

The identification of the same success pattern endows the intentional explanation with three different but related explanatory powers – generalization, predictive and counterfactuality powers. Let's assess them in detail.

Generalization, predictive and counterfactuality explanatory powers

Perhaps the *generalization power* is the most evident. The recognition of an intentional success pattern gives rise to a general explanation of different instances of the same success pattern in behavioural outputs triggered by different input stimuli. By contrast, non-intentional success explanations are capable of explaining why a specific behaviour achieved a given result, but not of providing a general explanation of why this behaviour originated by different input stimuli achieved the same result. This is a general and unified explanation that embraces all of these behavioural outputs. What makes the difference is that the positing of a representational state bridges across different input stimuli and hence give rises to intentional explanations of the form "organism *O* has the behaviour *B* because it represents *R*". The representation plays a unification role that makes viable the generalization over different behaviours triggered by different stimuli throughout the organism's history as well as present behavioural situations.

Suppose that an organism manifested a successful avoidance behaviour in different situations throughout its behavioural history, and that now it is also having an avoidance behaviour, and that what is common in all of these situations is the presence of a predator around. The pursued result is the avoidance of predators and hence there is a success pattern present in all of these cases. How do intentional and non-intentional success explanations explain the organism's avoidance behaviour? In what the difference consists between the intentional success pattern explained by the intentional explanation and the non-intentional success pattern explained by the non-intentional success explanation? The non-intentional explanation consists in the specification of a distinct and independent causal chain of a given behavioural output and of the conditions under which this behaviour succeeds in avoiding predators. But it does not establish any connection between this behaviour and other avoidance behaviours triggered by different stimuli, since it does not establish any connection between

their respective causal chains. The non-intentional explanation treats them as completely independent chains.

By contrast, the intentional explanation consists in a general explanation of the organism's behaviour in all situations: the organism represented the presence of the predator that triggered the avoidance behaviour which finally led to the successful behaviour of avoiding the predator. That is only possible because it abstracts from the specificities of the different stimuli that in each occasion triggered the behavioural output (e.g., a light array coming from certain direction, another light array with a different frequency, etc.) and instead unifies them by establishing what is common to all of them, namely, the presence of the predator. Hence, the intentional explanation is more powerful because it reaches a generality level that non-intentional success explanations are simply not capable of reaching. There are only individual and independent non-intentional explanations of specific behaviours in specific situations; there is nothing capable of unifying them by appealing to something in common throughout these specific behaviours. By contrast, in the case of intentional explanations, the positing of the representation of the predator is precisely the unifying element that provides the intentional explanations with the generalization power.

The positing of representational states also provides intentional explanations with *predictive power*. The identification of the intentional success pattern gives rise to the prediction that the organism with the token representation will behave in the same way and with the same success conditions that it behaved in past situations in which the pattern was present.[94] The representation is a predictive element that allows the intentional explanation to imply that the organism will have the same behaviour with the same success conditions, just

[94] There is no guarantee that because the organism represents the presence of the predator, then it will always have an avoidance behaviour. The organism may fail to have it because of an external problem (e.g., there is something that prevents the escape) or an internal problem (e.g., biological malfunctioning). Nevertheless, what matters is that the tokening of the representation of the predator will trigger an avoidance behaviour. In this example, I am just assuming that it will occur because there is no relevant external or internal problem that prevents the avoidance behaviour.

as it did in the past, provided that the same intentional success pattern is present. By contrast, the non-intentional success explanation is deprived of this predictive power because it fails to identify the intentional success pattern that is present despite the variety of input stimuli. That is the case because the non-intentional success explanation lacks an element that bridges across distinct stimuli that triggers the behavioural outputs – different input stimuli gives rise to distinct causal chains. In the absence of this unifying element, the intentional success pattern is not identifiable and hence the non-intentional explanation is not endowed with predictive power.

The former example illustrates quite well the predictive power. An intentional explanation of the organism's avoidance behaviour has the predictive power of establishing that the organism will escape whenever the intentional success pattern is present in the avoidance behaviour. The presence of this pattern entails that the organism represents the presence of the predator in the surrounding environment. The representational state provides the intentional explanation with such predictive power, it is the basis for the prediction that the representation of the predator will trigger the avoidance behaviour that will result in the escaping from the predator. By contrast, the non-intentional success explanation cannot make any prediction that the organism will have the same avoidance behaviour across different input stimuli because it fails to identify the pattern that bridges across these inputs. At best, it predicts what will happen in situations where the same stimulus input triggers the behavioural output (e.g., same position and orientation of the organism and predator, distance, etc.). What makes the difference is that in the case of the intentional explanation, future situations may involve very different input stimuli and still it predicts that the organism will have an avoidance behaviour provided that the intentional success pattern is present.

The difference between the predictive and generalization powers of intentional explanations lies in the distinct kinds of explananda. The predictive power has future

behavioural situations as explananda; the explanation consists in how the organism will behave in instances of a certain behavioural situation in the future. The generalization power has past and present behavioural situations as explananda; the explanation consists of a unified and general explanation of how the organism behaved in the past and behaves in the present in certain behavioural situations. In the case of the generalization power, the representation give rise to a general explanation of how the organism behaved in the past and behaves in the present, while in the case of the predictive power, the representation is an element that gives rise to an explanation of how the organism will behave in the future.

Finally, the last distinctive explanatory power of intentional explanations is the *counterfactuality power*. The identification of the intentional success pattern via the positing of the representational state allows the intentional explanation to explain what would have happened in counterfactual scenarios, provided that the intentional success pattern is present in the organism's behaviour. Suppose that in some past situation there was a predator in the surrounding environment and the organism had an avoidance behaviour. Now consider counterfactual scenarios in which things are different. For instance, the organism or the predator are in different positions or orientations; the surrounding environment has different weather or lighting, etc. What would then have happened in these counterfactual scenarios? According to the intentional explanation, the organism would have an avoidance behaviour because it represents the presence of the predator. What happens is that in these counterfactual situations, the behavioural outputs that pursue the avoidance of predators are triggered by different input stimuli, but the positing of the representational state allows the identification of the intentional success pattern despite such stimuli multiplicity. By contrast, the non-intentional success explanation would not be able to explain what would have then happened because these counterfactual scenarios involve behavioural outputs that are triggered by distinct input stimuli and hence by distinct causal chains. The non-intentional explanation lacks

the unifying element that would allow it to bridge over the variety of stimuli that triggered the causal chains that would lead to the successful behavioural outputs. Once again, the unifying element in the intentional explanation is the representational state.[95]

But what is the difference between the counterfactuality power of intentional explanations and the predictive and generalization powers? As previously noted, the fundamental distinction between the predictive and generalization powers is that the predictive power has as its explananda certain future behavioural situations, while the generalization power has as its explananda certain past and present behavioural situations. Hence, these explananda have in common that they are situations in the actual world, differing only on the temporal instances in which they occur. By contrast, the explananda of the counterfactuality power is not actual behavioural situations, but counterfactual behavioural situations.

At this point the following objection arises against my argument that the identification of the intentional success pattern endows the intentional explanation with predictive and counterfactual explanatory powers: non-intentional explanations support the prediction of what will happen in future scenarios and what would have happened in counterfactual scenarios. They are based on natural laws that give support to predictions and counterfactual assessments. Consider the natural law of the universal gravitation. Based on this, you can infer that because of gravity, if I had dropped the apple that is now in my hand, it would have fallen and reached the floor. Similarly, if the apple will be dropped, then it will fall. Hence, why is there no non-intentional explanation of counterfactual or future behavioural situations?

That is all true. However, once again the touchstone lies in the fact that the positing of the representational state allows the identification of the same success pattern throughout the organism's behavioural outputs, despite the variety of input stimuli which triggered it. The

[95] As previously showed in the third chapter, the counterfactual robustness argument also defends that counterfactuality power is a distinctive explanatory power of intentional explanations. Cf. STERELNY, 1995.

fundamental explanatory difference lies in the fact that the non-intentional success explanation is capable of explaining future and counterfactual scenarios that involve the same stimulus input. So, the non-intentional explanation is capable of explaining what will happen if the same stimulus occurs ("if the same stimulus will happen, the organism will have an avoidance behaviour") and what would have happened if the same stimulus would have occurred ("if the same stimulus would have occurred, the organism would have an avoidance behaviour"). By contrast, the intentional explanation is able to predict what will happen and to establish what would have happened in scenarios in which a variety of input stimuli will (or would have occurred) precisely because it is capable of bridging across these distinct stimuli in virtue of the identification of the same success pattern throughout all of these scenarios. The positing of the representational state is what makes it possible for the intentional explanation to identify the same success pattern and hence to bridge across these distinct stimuli. In sum, the intentional explanation has more powerful counterfactual and predictive explanatory powers because the amount of counterfactual and future scenarios that it is capable of explaining is much bigger than the amount of counterfactual and future scenarios that non-intentional explanations are capable of explaining.

Representational systems

So far, so good. I have defended that the presence of a success pattern on the behaviour of a given organism is a minimal condition for it to be an intentional system. However, the mere presence of a success pattern is not enough for minimal intentionality because there are intentional and non-intentional success patterns. The fundamental distinction between them is that the intentional one bridges across a variety of input stimuli while the non-intentional one fails to do that – the behavioural output is always triggered by the same type of stimulus. This is a variation of the constancy mechanism condition because it is applied for the recognition of

the same success pattern, despite the multiplicity of input stimuli. Based on this constancy mechanism condition, it becomes clear that success patterns present in the behavioural output of several very simple organisms are not intentional, while the success pattern of other organisms are intentional. Let's consider some illustrative cases.

Consider the osmolarity example again. The hypothalamus produces the antidiuretic hormone to lower the osmolarity of the human blood when it goes above a certain level. It is clear that there is a success pattern present here since the hypothalamus has a pursued result – to keep the osmolarity at a certain level – and produces the antidiuretic hormone to reach that result – there will be success whenever this result is achieved. Furthermore, the kidney uses the hormone as a proxy for the high osmolarity to lower the osmolarity in the blood. Nevertheless, the mere presence of such success pattern is not enough for the conclusion that the hypothalamus is an intentional system. This is the case because the production of the antidiuretic hormone by the proper functioning hypothalamus is only triggered by the same kind of stimulus – high osmolarity.[96] The hypothalamus does not bridge across a variety of stimuli in the production of the antidiuretic hormone in order to achieve the biological goal of keeping the osmolarity at a certain level. Thus, the success pattern present in the osmolarity example is not intentional.

I have introduced the distinction between intentional and non-intentional success patterns by contrasting intentional systems that are able to bridge across a variety of input stimuli and non-intentional systems that fail to do that. Let's illustrate the tropistic systems with the paramecium case. Whenever light is present, the paramecium goes in the opposite direction. Its behavioural output is always triggered by the same kind of stimulus, light. Its success pattern is constituted by the paramecium's biological goal that is achieved by escaping

[96] The specification that high osmolarity is the only type of stimulus that triggers the production of the hormone in the proper functioning hypothalamus is required because, when malfunctioning, the hypothalamus produces the hormone even when the osmolarity is not high and so will be triggered by other stimuli.

from light. But it is not an intentional success pattern since the non-intentional explanation has the same explanatory power as the intentional explanation precisely because the behavioural output does not bridge across a variety of input stimuli. Rather, it is only triggered by the same light stimulus. For the same line of reasoning, the anaerobic bacterium is not an intentional system. There is a clear success pattern in its behavioural output constituted by the biological goal of reaching oxygen-free water (oxygenated water is fatal to it) via the use of internal magnetesomes (the direction of the geomagnetic field coincides with the direction of the bottom of the ocean that is oxygen-free). However, the bacterium's behaviour is only triggered by one kind of stimulus – the prevailing magnetic field – and so does not bridge across a variety of stimuli to produce the behaviour of moving to the oxygen-free water. In sum, both paramecia and bacteria are not intentional because they fail to satisfy the constancy mechanism condition.

What about organisms that satisfy the success pattern and constancy mechanism conditions and hence are intentional according to the dual proposal? There are plenty of cases. Let's start with the less simple case and then pass to the simpler one. Vervet monkeys in Kenya have a complex signalling system. As Seyfarth, Cheney and Marler (SEYFARTH et al., 1980) present in their classic paper, vervet monkeys give acoustically different alarm calls to at least three different predators: leopards, martial eagles and pythons. The alarm signals are produced by the speaker monkey to signal to other monkeys the presence of predators, so they can have specific avoidance behaviours that vary from signal to signal. Let's concentrate on eagle alarms. When monkeys are on the ground, eagle alarms cause them to look up and/or run into trees in order to avoid the eagle's stoops. When in the trees, eagle alarms cause them to evoke looking ups and/or running out of the tree. The authors conclude that these distinct responses to the alarms suggest that "each alarm call effectively represented, or signified, a different class of external danger" (SEYFARTH et al., p. 802), that is, the eagle's presence. These responses also occur in the absence of alarms, but because the other monkeys actually see the predator.

So, according to the dual proposal, do vervet monkeys represent the presence of eagles? There is a clear success pattern constituted by the alarming call and the subsequent avoidance behaviour. The relevant state also bridges across different proximal stimuli to represent the same external feature – the eagle. It is perspicuous that the monkey tokens this state and prompts the avoidance behaviour at different times, positions, distances, and angles, which implies the employment of a constancy mechanism in the production of the state and so that it bridges across a variety of proximal stimuli. For instance, difference in environmental light in virtue of the different sun positions in the sky. The same line of reasoning shows that vervet monkeys also represent leopards and pythons.[97]

Even though vervet monkeys are not as complex organisms as chimps or baboons, they are still complex. It may appear then that the dual proposal is too restrictive – it does not treat certain states as representations when they clearly are. So, let's show that the dual proposal treats even simple organisms like certain insects as intentional. Honeybees perform waggle dances whose properties correlate with the distance from the hive of some sources of nectar. They dance in order to show other honeybees the direction of the source of nectar.[98] Does the honeybee represent the source of nectar? Once again, the waggle dance and the triggering of the foraging behaviour have a success pattern – to gather nutrients – and there is the use of the waggle dance as a proxy for nectar. It is also clear that the honeybee represents the same source of nectar despite the variety of proximal stimuli. Because the flowers where the honeybee gets nectar are likely to change every few days when they are in bloom, the honeybee evolved the capacity to learn colours and shapes accurately. For instance, despite the change of colour or

[97] There is strong evidence that the vervet monkeys perceptually categorizes the represented predators: "By giving alarm calls to some species but not to others, and by giving acoustically distinct alarms to different predators, vervet monkeys effectively categorize other species. More than 1000 species of mammals, birds and reptiles were seen regularly by the monkeys without eliciting alarm calls" (SEYFARTH et al., 1980, p. 803).

[98] Karl von Frisch is famous for his discovery of the honeybees' communication system (VON FRISCH, 1967). I have collected the following information on the cognitive capacities of honeybees from Mandyam Srinivasan's paper "Honey Bees as Model for Vision, Perception and Cognition" (SRINIVASAN, 2010).

shape of a given flower, the honeybee can still represent it as a source of nectar. They "possess color constancy, which is the ability to determine the true color of an object […] independently of the spectrum of the illumination under which it is viewed" (SRINIVASAN, 2010, p. 279). Such capacity enables the honeybee to accurately identify the colour of the flower whether it is in the sun or under the shade of tree. The honeybee employs a constancy mechanism in the production of the representational state of the location of nectar, otherwise it would not represent the same flower as a source of nectar despite variations of colour and shape.[99] So, honeybees are intentional systems.

Here it may be objected that it is not the case that the honeybee satisfies the constancy mechanism condition. I have argued that the honeybees satisfy this condition, in opposition to the hypothalamus in the osmolarity example. But here it could be objected that there is a variety of proximal stimuli in both the honeybee and hypothalamus cases. The reason is that the production of the antidiuretic hormone by the hypothalamus is triggered by a variety of (high) osmolarity levels in the blood, just like the honeybee representation of the flower is triggered by a variety of colours and shapes. Hence, why say that only the honeybee employs the constancy mechanism? The fundamental distinction is that the production of the antidiuretic hormone by the hypothalamus is triggered by different degrees of the *same kind* of stimulus (different levels of osmolarity in the blood, but always above a certain level), while the honeybee's representation is triggered by *different kinds of stimuli* – different colours and shapes of the flower. It is not that the honeybee's representation is triggered by just different degrees of a given colour or shape, but that it is triggered by different colours and shapes. It is because of this distinction that the honeybee employs the constancy mechanism, but not the hypothalamus.

[99] More evidence that honeybees employ constancy mechanisms is that they can represent the same external feature despite the fact that it is poorly visible or camouflaged, cf. SRINIVASAN, 2010, p. 274.

Input and output explanatory powers

I have proposed that it is a minimal condition for a state to represent an external feature that it bridges across a variety of proximal stimuli coming from the feature. Does this bridging condition also apply to the behavioural output? That is, is it a condition for intentionality that the organism should produce a variety of behavioural outputs to achieve the pursued result for it to be intentional? I think that no such minimal condition is required. Provided that the organism produces the behavioural output to achieve its pursued result, it is not required that the organism should produce a variety of behavioural outputs in order for it to be intentional. So, room is open for the organism to be intentional even if it always has the same behavioural output triggered by the tokening of the supposed representational state.

But isn't there an analogous explanatory argument in favour of this requirement? Let me explain. I have defended the requirement that intentional success patterns bridge across a variety of input stimuli by appealing to the distinctive explanatory powers of intentional explanations. Intentional success patterns should have a variety of input stimuli because otherwise there would be no room for the distinctive explanatory powers of intentional explanations, i.e., counterfactuality, predictive and generality powers. So why not appeal to the analogous explanatory argument according to which intentional success patterns should bridge across a variety of behavioural outputs because otherwise there would be no room for the distinctive explanatory powers of intentional explanation? It could be argued that the bridging across a variety of input stimuli by the intentional success pattern is not enough to keep room open for the distinctive explanatory powers of intentional explanations – it is also required that the intentional success pattern also bridges across a variety of behavioural outputs. What is wrong with this argument? I think that the requirement that the intentional success pattern

bridges across a variety of input stimuli is enough to explain why intentional explanations have predictive, generality and counterfactual powers. The contrast with the requirement that the intentional success pattern should also bridge across a variety of behavioural outputs is clear. Let me show why.

The positing of the relevant internal state as representational is justified because otherwise the recognition of the same success pattern as being present despite the variety of input stimuli would not be possible. Notice that the positing of the internal state is the unifying element for the recognition of the presence of the same success pattern despite the distinct input stimuli. In the absence of this state playing this unifying role, the recognition of the same success pattern would not be possible. By contrast, the positing of the internal state as representational is not required for the recognition that the same pattern is present despite the variety of behavioural outputs. This is the case because no matter how different the behavioural outputs are, one can still recognise the presence of the same success pattern simply because they all are pursuing the same result. After all, they are produced by the organism in order to achieve the organism's pursued result. So, the pursuit of the same result is the unifying element for the recognition of the presence of the same success pattern despite the variety of behavioural outputs. But if that is the case, it follows that the positing of the internal state as a representation would make no explanatory difference for one to recognise the presence of the same success pattern despite the variety of behavioural outputs. That is, such positing would not play the unifying role required to recognise the same success pattern. Rather, the fact that the behavioural outputs pursue the same result is the one that plays this unifying role. Hence, the requirement that the representational state should bridge across a variety of behavioural outputs would deliver no explanatory purchase.

But how would one still know that the distinct behavioural outputs pursue the same result? Notice that this problem also applies to cases in which the internal state bridge across a

variety of input stimuli but not across a variety of behavioural outputs – it always triggers the same behavioural output. Even in these cases, it may be not clear which result the organism pursues via the production of the same behavioural output. The only way to determine what is the pursued result of the behavioural output(s) of a given representational state is by proposing a criterion that specifies the pursued results of representational states. For instance, according to teleosemantics the pursued result is the biological function of the representational state. The problem of the specification of the pursued result of distinct behavioural outputs is just an instance of the general problem of the specification of the pursued result of a representational state, whether it triggers a variety of behavioural outputs to achieve this result or whether it triggers always the same behavioural output. But if what determines the pursued result of the representational state is the adopted criterion (e.g., the biological function of the representational state), then it does not matter whether the state triggers a variety of behavioural outputs or not. The conclusion is that the requirement that the representational state should bridge across a variety of behavioural outputs is not a minimal condition for intentionality.

I have developed the dual proposal for minimal conditions for intentionality constituted by the success pattern and the constancy mechanism conditions. According to the first, it is a minimal condition for a given state to be representational that (i) a success pattern is present in the behavioural output triggered by the internal state; and (ii) the system use the state as a proxy for the presence of the relevant external condition. According to the constancy mechanism condition, the distinction between an intentional and nonintentional success pattern is that the internal state bridges across a variety of input stimuli coming from the relevant external feature. In both cases, the defence was based on the explanatory powers given rise to respectively by the presence of the success pattern in the behavioural output and by the bridging across a variety of input stimuli by the representation. But what is the relation between these two explanatory powers? What distinguishes one from the other? On the one hand, if there is no relevant

distinction among them, the establishment of one of them is superfluous. On the other hand, if there is a relevant distinction, there is always the threat that they are incompatible.

Let's call "success explanatory power" the explanatory power given rise to by the presence of the success pattern on the behavioural output and "constancy explanatory power" the explanatory power given rise to by the internal state's bridging across a variety of proximal stimuli coming from the relevant external feature. The fundamental distinction between the success and the constancy explanatory powers is that the first originates at the output process that is triggered by the tokening of the representation, while the second originates at the input process that causes the tokening of the representation. That is, the success explanatory power comes from the input side while the constancy explanatory power comes from the output side. Let me explain. The presence of the success pattern is required in the behavioural output triggered by the representational state. There is a success pattern provided that the internal state pursues a given result by triggering a behavioural output to achieve it. Hence, the presence of the success pattern gives rise to an explanatory power that is generated by a constraint on the relevant behavioural output process. Namely, the behavioural output that is triggered by the tokening of the internal state. In sum, the success explanatory power is an *output explanatory power*. On the other side, the employment of the constancy mechanism is a requirement that the internal state should be able to still represent the same external feature despite the variety of stimuli coming from it. So, whatever the explanatory power given rise to by the bridging across a variety of input stimuli, it originates with a constraint on the input process that triggers the tokening of the internal state. In sum, the constancy minimal condition is an *input explanatory power*. This is the reason that the dual proposal is not *ad hoc* or arbitrary. It is not a proposal developed only to neutralize the objection of liberality or a proposal that arises from an arbitrary conjunction of two minimal conditions that give rise to two different and unrelated explanatory powers. Rather, it is a proposal that naturally arises from the verification that the

constancy mechanism condition originates from a constraint on the input process that triggers the tokening of the internal state and that the success pattern condition originates from a constraint on the output process that is triggered by the tokening of the internal state.

What about the similarities? Remember that the explanatory power given rise to by the presence of the success pattern is the power of explaining success. It specifies the external feature under which the behavioural output triggered by the internal state achieves the pursued goal – the presence of the feature guarantees the success of the behavioural output. The positing of a result pursued by the internal state also gives rise to the generality, predictive and counterfactual explanatory powers, just as the explanatory power given rise to by the employment of the constancy mechanism. The positing of the pursued result allows the recognition that the same success pattern is present despite the variety of behavioural outputs. This recognition gives rise to the generality power because it provides a general and unified explanation of different behavioural outputs, produced in different instances, pursuing the same result. The explanation bridges across different behavioural outputs by recognising the presence of the same success pattern in these different situations. The positing of the external result pursued by the internal state also gives rise to predictive and counterfactuality powers. It provides an explanation with these explanatory powers because it affords the recognition that, despite the production of different behavioural outputs in future (or counterfactual) situations, the organism will (or would) pursue the same result.

The success pattern condition does not require that the organism should be able to produce a variety of behavioural outputs to achieve the pursued result. As long as the organism has a success pattern in its behavioural output and uses the internal state as a proxy for the presence of the relevant external condition, the success pattern condition is satisfied. So, room is open for the internal state to always trigger the same behavioural output and still satisfy the success pattern condition. In these cases, the presence of the success pattern would not give

rise to the generality, predictive and counterfactuality powers since there would be no variety of behavioural outputs. After all, if an organism always produces the same behaviour triggered by the internal state in response to the external feature, then the positing of the external result pursued by this state does not make any difference for an explanation of general cases or for what will or would happen. So, there is only the unique explanation that the organism can produce only one behavioural output to achieve the relevant result. Nevertheless, such positing would still give rise to the explanatory power of explaining success. That is, the explanation specifies the external condition under which the behavioural output triggered by the internal state succeeds in achieving the pursued result.[100] The conclusion is that the explanatory power of the success pattern condition is guaranteed both in situations where there is a variety of behavioural outputs as well as when the behavioural output is uniform.

Minimal distance and the indeterminacy objection

The dual proposal here developed assumes the constancy mechanism condition as a minimal condition for intentionality. However, in the third chapter I assessed and rejected the constancy mechanism proposal for the limits of intentionality. Both proposals assume that it is a minimal condition for intentionality that the system should employ a constancy mechanism in the production of the state in order to bridge across a variety of input stimuli and still represent the same external feature. However, the constancy mechanism condition required in the dual proposal is a variation of the constancy mechanism proposal, since it is here applied to distinguish non-intentional from intentional success patterns. That is, it is applied to recognise the presence of the same success pattern in situations where there are a variety of

[100] Notice that the fact that the success pattern may be present in the behavioural output of organisms without a variety of behavioural outputs is another reason for the requirement of the constancy mechanism condition. In cases of uniformity of the behavioural output to achieve the pursued result, the satisfaction of the success pattern condition is not sufficient to justify the positing of the representational state from an explanatory point of view.

proximal stimuli. By contrast, there is no such application in the constancy mechanism proposal. This proposal is not committed to any success pattern. Indeed, Sterelny and Burge, the main proponents of the constancy mechanism proposal, firmly reject the appeal to success patterns to draw the limits of intentionality (cf. BURGE, 2010, pp. 292-308; STERELNY, 1995).

In the third chapter I developed a general objection to the constancy mechanism proposal – the indeterminacy objection. It maintains that since the distinction between proximal and distal features comes in degree, and that there is no non-arbitrary way of drawing a line to strictly divide proximal from distal features, the following problem arises for the constancy mechanism condition: what is the minimal distance from the organism's sensory apparatus that the stimulus should be for the internal state to be genuinely intentional? Note that for a state to bridge across proximal stimuli in order to represent the same external feature, *some* distance between the organism and the represented external feature is required. Otherwise there would be no distance based on which the constancy mechanism would keep the state representing the same external feature despite the variety of proximal stimuli. But what is the extent of this minimal distance?

I concluded that the extent of this minimal distance is indeterminate in light of the constancy mechanism proposal, since it does not provide a non-arbitrary criterion to determine it. But things change when this problem is considered in light of the dual proposal that holds the constancy mechanism condition together with the success pattern condition. The success pattern condition gives rise to a straightforward and non-arbitrary criterion to solve this problem. It is perspicuous that the presence of the success pattern in the behavioural output presupposes that the organism pursues a given result towards a certain feature of the external environment (e.g., an organism that tries to escape from predators). Evidently, there is a certain distance between this external feature and the organism's sensory apparatus. But if the distance

is so short that it precludes the presence of the success pattern on the behavioural output triggered by the state, then it follows that this state is not intentional. So, how to establish how short this distance may be?

This is the criterion: the minimal distance is the one that does not preclude the presence of the success pattern on the behavioural output triggered by the internal state. Provided that this criterion is satisfied and so the presence of the success pattern in the behavioural output is not precluded, it does not matter how short the distance is. For instance, suppose that a stimulus chain triggers the tokening of the internal state in a certain organism – s^1 is the most proximal stimulus in the chain, s^2 is the second most proximal, s^3 is the third one etc. Evidently, the internal state cannot represent s^1 because there is no distance between the sensory apparatus and the stimulus. May it represent s^2? It is plainly possible, provided that this short distance keeps room open for the presence of the success pattern on the organism's behavioural output. If that is indeed the case on concrete cases is a further question that depends on the specific features of each concrete case. Finally, notice that even though the distance to s^2 is short, at this stimulus level the employment of the constancy mechanism still gives rise to the counterfactuality, generality and predictive explanatory powers and hence to a relevant explanatory power.

Here it is necessary to make an observation on the counterfactual robustness argument for the constancy mechanism proposal. In the previous chapter, I have rejected it by appealing to the fact that at the second most proximal stimulus level the group of counterfactual situations would be so small that it would not give rise to a relevant counterfactual explanatory power. In the end, this line of reasoning leads to the conclusion that it is indeterminate what is the first stimulus in the causal chain for which the positing of the state as representing this first stimulus gives rise to a robust explanation. The indeterminacy objection threatens the counterfactual robustness argument because this argument fails to notice that the employment of the constancy

mechanism gives rise not only to the counterfactuality explanatory power, *but also to the generality and predictive powers*. So, even at the second most proximal stimulus level the employment of the constancy mechanism gives rise to a relevant explanatory power constituted by these three explanatory powers. The conclusion is that the dual proposal is immune to the indeterminacy objection because it claims that the employment of the constancy mechanism gives rise to these three explanatory powers.

It should be highlighted that, as previously argued, the presence of the success pattern in the organism's behavioural output gives rise to an explanatory power of the intentional explanation. The above criterion that the minimal distance is the one that keeps room open for the presence of the success pattern guarantees this explanatory power. After all, if the distance is so short that it precludes the presence of the success pattern, there is no place for this explanatory power to arise. As previously shown, the success explanatory power and the constancy mechanism explanatory power are perfectly compatible and complementary. This criterion for the minimal distance makes sure that the fulfilment of one condition – the constancy mechanism condition – does not preclude the fulfilment of the other condition – the success pattern condition.

Here the following objection may arise. If this criterion opens the possibility for the minimal distance to be the second proximal stimulus to solve the minimal distance problem, why can't the proponents of the constancy mechanism proposal claim that the second proximal stimulus is the minimal distance? If the criterion defended here works well for the dual proposal, why it can't also work for the constancy mechanism proposal? What is problematic is not the adoption of this or that criterion to determinate the minimal distance, but rather how this criterion is principled justified in accordance with the constancy mechanism proposal to avoid arbitrariness. It is hard to conceive how it can be done since this proposal is solely constituted by the constancy mechanism condition, which implies that the justification of this

criterion would have to be solely based on this condition. Otherwise, the justification would be plainly *ad hoc*. If the stimuli variety given rise by the second proximal stimulus criterion constitutes the "sufficient variation" (STERELNY, 1995, pp. 261-2) for intentionality, what is the principled justification for it in light of the very constancy mechanism condition?

By contrast, the dual proposal is constituted by the constancy mechanism and success pattern conditions. It claims that the explanatory power given rise to by the satisfaction of both conditions constitutes *the distinctive* explanatory power of representational states. The criterion defended here that the minimal distance is the one that does not preclude the presence of the success pattern in the behavioural output is based on the explanatory power given rise to by these conditions. It may turn out that this criterion may be satisfied by the second proximal stimulus in some concrete cases, but here this minimal distance is justified. It is not arbitrary. Together with the explanatory power given rise to by the satisfaction of the success pattern condition, even the explanatory power of small varieties given rise to by the minimal distance of the second proximal stimulus justifies the positing of the representation. The non-intentional explanation would lack such explanatory power. This explanatory justification, however, is not allowed for the proponent of the constancy mechanism proposal. For them, the only relevant explanatory power is the one given rise to by the varieties of proximal stimuli in virtue of the employment of the constancy mechanism. That is the reason the constancy mechanism condition cannot stand by itself. United with the success pattern condition it stands, divided it falls.

But why can't proponents of the constancy mechanism proposal also say that the explanatory power given rise by the second most proximal stimulus distance is enough for the positing of the representational state to be explanatorily justified? Notice that they require sufficient variation for intentionality. They could then argue that the explanatory power given rise by the second most proximal stimulus distance is enough variation. Maybe such variation

is enough for the explanatory power of the resulting intentional explanation to trump the explanatory power of non-intentional explanations. However, such a move is problematic. In the dual proposal's case, it is easy to show that the explanatory power given rise to by the second most proximal stimulus distance is enough variation since it appeals to the explanatory power given rise to by both the constancy mechanism and the success pattern conditions. By contrast, the constancy mechanism proposal can appeal only to the explanatory power given rise to by the employment of the constancy mechanism. But why does it constitute enough variation? This is not clear at all. The conclusion is that until some convincing justification is provided, the constancy mechanism proposal is flawed in light of the indeterminacy objection.

Let's finish this section with a summary of the dual proposal for the minimal conditions for intentionality. I started the investigation for the explanatory power of representational states with the success pattern condition by arguing that the presence of the success pattern gives rise to the success explanatory power. However, this condition is not enough for intentionality because the resulting theory would be too liberal. The non-intentional explanation would have the same success explanatory power by specifying the causal chain and the result pursued by the relevant state without positing any representational state. Thus, I have argued that another minimal condition is required to distinguish intentional from non-intentional success patterns – the constancy mechanism condition. The distinction is that the positing of the representational state provides the intentional explanation with generality, predictive and counterfactuality powers because the representational state bridges across a variety of proximal stimuli and still represents the same external feature. After that, I showed that the success pattern and the constancy mechanism explanatory powers are compatible and complementary because the first is an output explanatory power and the second is an input explanatory power, which shows that the dual proposal is not *ad hoc* or arbitrary. Finally, I showed that the minimal distance problem that threatens the constancy mechanism proposal rejected in the previous chapter does not

threaten the dual proposal. The reason is that the latter proposal gives rise to the criterion that the minimal distance is the one that does not preclude the presence of the success pattern on the behaviour triggered by the representation.

In sum, the dual proposal maintains that the explanatory power of intentional explanations is constituted by the success and constancy explanatory powers. The satisfaction of the success pattern and constancy mechanism conditions guarantees that the positing of the relevant representation plays the distinctive explanatory role of representational states. But is the dual proposal compatible with the intuitive conception of representation? If so, what is the degree of such compatibility? Let's move to the intuitive side of the debate on the problem of demarcation.

4.4 Is the dual proposal intuitive?

In the last chapter and in the previous sections of this chapter, I made an extensive investigation focused on the explanatory role played by representational states in intentional explanations of behaviour. I assessed and rejected the causal independence proposal developed by Fodor and Beckerman (section 3.4) and the constancy mechanism proposal developed by Kim Sterelny and Tyler Burge (section 3.5). After that, in this chapter, I developed the dual proposal that establishes that the positing of a representational state by an intentional explanation is explanatorily justified only if it satisfies the success pattern condition and the constancy mechanism condition. I have also showed that the causal independence and the constancy mechanism proposals violate highly entrenched intuitive principles in our intuitive conception of representation. Now let's focus on the intuitiveness of the dual proposal.

Is the dual proposal defensible in light of the intuitive conception of representation or is it so counter-intuitive that it should be ruled out? Finally, how should we balance, in the context of mutual adjustments, the explanatory power of positing a given representational state

established by the dual proposal with the counter-intuitive aspects of this proposal? In this final section, I claim that the dual proposal is compatible with the intuitive conception of representation to an acceptable extent by arguing that it is fully compatible with highly entrenched intuitive principles. Beyond this, the limits of intentionality become blurry and it turns to be a pragmatic choice to accept or reject the intentional status of a given state. However, such blurriness is not problematic, especially for naturalist theories of representation like teleosemantics. I will start the intuitive assessment of the dual proposal by recapitulating the intuitive principles that I defended in the last chapter as highly entrenched in our intuitive conception of representation.

First, there is the autonomy principle. Intentional systems are autonomous: the forces responsible for their behavioural outputs originate within the systems, not outside. They are self-moving systems. So, it is required for intentionality that the system's behavioural output is not (wholly) caused by forces that originate outside the system. Second, there is the complexity principle. Systems without a certain cognitive complexity are not genuinely intentional; it is implausible to claim that very simple systems like viruses or unicellular organisms are intentional because they lack the minimal cognitive complexity required for intentionality. The third and last intuitive principle is the activity principle. It requires that, for a given system to be intentional, it should have some active role, i.e., it cannot be wholly passive. A system is not intentional in cases when the internal state is automatically triggered by the presence of the relevant external feature and the state always triggers the behavioural output. The passivity arises from the fact that the behavioural output is automatically activated by the relevant specific stimuli. Thus, the activity principle rules out passive systems as genuine intentional systems.

Let's start with the complexity principle. Is it compatible with the success pattern and constancy mechanism conditions established by the dual proposal? As I have previously argued

in the last chapter, the requirement that a system should employ a constancy mechanism in the production of representational states implies that the system should have a certain level of cognitive complexity. After all, the capacity to employ this mechanism makes the relevant system more complex than it would be in the absence of such capacity. Very simple systems (paramecia, hypothalami, etc.) are ruled out as representational systems, in opposition to systems that employ constancy mechanisms (honeybees, vervet monkeys etc.). On the other hand, the success pattern condition seems to be incompatible with the complexity principle, since systems as simple as hypothalami satisfy this requirement. However, what is relevant here is not whether, in isolation, the success pattern condition is compatible with the intuitive conception of representation. What matters is whether the global picture of intentionality delivered by the dual proposal is compatible with this intuitive conception. Since this proposal is also constituted by the constancy mechanism condition, it follows that the dual proposal requires a certain complexity from the relevant system for it to be intentional.

What about the autonomy principle? The success pattern condition establishes that it is a minimal condition for intentionality that the system has a success pattern in the behavioural output triggered by the representational state, and that the system uses it as a proxy for the presence of the relevant external feature in the production of the behavioural output. This condition rules out automaton systems. After all, if the behavioural output is fully generated by forces originating outside of the system, it follows that there is no use of the state as a proxy for the presence of the external feature and hence that the success pattern condition is not satisfied. On the other hand, as I have shown in the last chapter, the satisfaction of the constancy mechanism by a given system entails that it is autonomous. The employment of the constancy mechanism requires a selection process of picking from all input stimuli just the ones that correlate with the external feature, and such a selection process is an internal force that affects the resulting behavioural output. In conclusion, the success pattern and constancy mechanism

conditions are both compatible with the autonomy principle.

Now consider an autonomous system that has a certain cognitive complexity and the forces responsible for its behaviour are internal. So, it respects the autonomy and complexity principles. Nevertheless, the system is deprived of any active role. The relevant state is automatically triggered by the presence of the external feature and this state always triggers the behavioural output. That is, the behaviour is automatically activated by this specific stimulus. Now, contrast again tropistic systems with systems that satisfy the constancy mechanism condition. As previously shown, the employment of the constancy mechanism does not guarantee that the system plays some relevant active role. However, the satisfaction of the success pattern condition requires an active role from the system. To use the state as a proxy for the presence of the external feature in the production of the behavioural output is an active role played by the system. After all, to use something (the state) as a guide for producing something else (the behaviour) is an activity – it is not possible for a passive system to *use* anything at all.[101]

The general picture is that the satisfaction of the constancy mechanism condition guarantees that the organism respects the complexity principle; the satisfaction of the success pattern condition guarantees that the organism respects the activity principle; finally, the satisfaction of the constancy mechanism or the success pattern conditions guarantees that the organism respects the autonomy principle.

As argued in the last chapter, I think that these three principles are the most entrenched in our intuitive conception of representation. Since the dual proposal respects them, it seems that it is compatible with the intuitive conception of representation. But what about other principles? This is not an exhaustive list of intuitive principles of representationality. Evidently,

[101] I have claimed that the satisfaction of the success pattern condition does not require that the system is able to produce several behavioural outputs. But how can a system that produces the same behaviour have some active role? Once again, because the system uses the internal state to achieve a given result by producing this behaviour. The use of the state makes the system active even though it always produces the same behaviour.

it is possible to consider another supposed intuitive principle that is incompatible with the dual proposal because it rules out as representational certain states that satisfy both the success pattern and the constancy mechanism conditions. Let's suppose for the sake of the argument that there is indeed such intuitive principle. So, should we immediately get rid of the dual proposal? I don't think so.

The dual proposal is supported on the one hand by the explanatory power given rise to by the success pattern and constancy mechanism conditions, and on the other hand by its compatibility with the complexity, autonomy and activity intuitive principles. In order for such an intuitive principle to justify the rejection of the dual proposal, it should override the explanatory and intuitive reasons that support the dual proposal. That is, it should be a very entrenched and very intuitive principle to have such radical consequence. It is not plausible to suppose that there is an intuitive principle or set of intuitive principles so strong.

But now suppose that the dual proposal treats certain states as representational, even though it is counter-intuitive, in relevant aspects, to claim that they are genuine representational states. It is very plausible to suppose that there are such cases. Once again, this is not a reason strong enough to justify the rejection of the dual proposal. However, it represents a more serious problem for this proposal because it is more plausible to suppose that there are indeed such cases. So, is some further reason required for the dual proposal to still stand as a viable proposal to demarcate the limits of intentionality?

I think that this is the point on the development of the debate on the limits of intentionality in which we have reached the *real* borderline cases between primitive representational and non-representational states. There will always be a stage of this debate in which we will find some candidates for representational states that, even though they satisfy the explanatory and intuitive requirements established by the dual proposal, it still looks really counter-intuitive to consider them as genuine representations. This is where the limits of

intentionality become blurry. Nevertheless, this is not problematic for the dual proposal.

Every minimally viable demarcation proposal will face cases in which the limits of intentionality that it draws becomes blurry. There will always be good and incompatible reasons for accepting, as well for rejecting, the representational status of candidates for representational states considering the very criterion established by the proposal. That is, there will be a good reason for claiming that a given state is representational, but also a good reason for claiming that is it not. This is the case because it is highly implausible for a demarcation proposal to draw fully strict limits of intentionality. That is, it is implausible that there is any proposal that demarcates a strict limit for intentionality such that (i) there are no good and incompatible reasons for accepting and rejecting the representational status of a given candidate state; and (ii) this proposal is still minimally viable in light of explanatory and intuitive considerations. Evidently, you can always develop a fully strict proposal for the limits of intentionality with no limiting cases. However, the price of such strictness is the loss of minimal explanatory or intuitive viability.[102] So, how to decide the representational status of these limiting cases?

I think that at this point it turns out to be just a pragmatic choice of accepting or rejecting the intentionality of limiting cases. You can either decide that a given state is not representational because it does not look intuitively representational or you can decide that it is representational because what matters is that it comes with enough explanatory purchase. The choice between these decisions is pragmatic, it varies in different contexts. This is precisely what happens in limiting cases.

Finally, the blurriness at this stage of the limits of intentionality drawn by the dual

[102] Consider the proposal that the only minimal condition for intentionality is that the candidate is an internal state of my organism. It is strict because for every possible candidate you can decide whether it is representational or not by just verifying whether or not it is in my organism. But this proposal is indefensible from both explanatory and intuitive perspectives. Even the more plausible proposal that the only representations are human internal states is not fully strict because there are still limiting cases. For instance, are some foetuses' internal states of representations given that it is contentious whether foetuses are humans?

proposal is not problematic for teleosemantics and other naturalist theories of mental representations because what threatens their viability in the debate on the limits of intentionality is the objection of liberalism. They are attacked because they treat certain states as representations when they are clearly not representational. However, as I argued in the last section, the constancy mechanism condition neutralizes this objection by establishing an explanatory requirement that makes the dual proposal much more restrictive. Notice that the objection of liberality is based on cases in which the naturalist theories treats certain states as representations when they are *clearly* not representational. But at the stage at which the limits of intentionality drawn by the dual proposal becomes blurry, the relevant candidates are not either clearly representational or not. So, the objection of liberality does not even arise.

Conclusion

The dual proposal has certain similarities with Schulte's demarcation proposal (SCHULTE, 2015). Schulte proposes the adoption of the constancy mechanism condition by teleosemantics to neutralize the objection of liberality. However, the proposals are fundamentally distinct. First, they diverge on the resulting demarcation of the limits of intentionality. Schulte's proposal is committed to the thesis that it is the biological function that determines the organism's pursued result, while the dual proposal remains neutral on that issue. The dual proposal just requires the existence of *some* pursued result, no matter whether it is functionally determined or not. Second and most important: the justifications of both proposals are fundamentally distinct. Schulte justifies his proposal by appealing to the counterfactual robustness argument. That argument, however, is flawed as previously demonstrated in the third chapter. The justification of the dual proposal is based on the method of reflective equilibrium which consist in mutual adjustments between explanatory and intuitive constraints. The explanatory justification of the success pattern condition is that its

satisfaction guarantees the explanation of success. By contrast, the satisfaction of the constancy mechanism condition distinguishes intentional and non-intentional success patterns. Its explanatory justification is that the employment of the constancy mechanism gives rise to generality, predictive and counterfactual explanatory powers by allowing one to recognise the presence of the same success pattern despite the variety of input stimuli. The fundamental distinction between the success and constancy explanatory powers is that the former originates at the output process triggered by the tokening of the representation, while the latter originates at the input process that causes the tokening of the representation.

In this chapter, I developed the dual proposal for the minimal conditions for intentionality constituted by the success pattern and constancy mechanism conditions. I have argued that the distinctive explanatory power of intentional explanations is constituted by the success pattern and the constancy explanatory powers. In the final section, I showed that the dual proposal is compatible with the intuitive conception of representation by arguing that it respects three highly entrenched intuitive principles – the complexity, autonomy, and activity principles. Thus, the conjunction of the success pattern and constancy mechanism conditions is justified in light of both explanatory and intuitive requirements. I take this to be the optimal state in the revisionary process of the method of reflective equilibrium. The success pattern and the constancy mechanism conditions should be revised no more. Therefore, the dual proposal draws the appropriate demarcation of the limits of intentionality and hence solves the problem of demarcation.

> **The dual proposal.** A given state represents an external feature only if it satisfies two minimal conditions for intentionality. (I) the success pattern condition – there is a success pattern in the system's behavioural output and the system uses the state as a proxy for the presence of the external feature in the production of such

behavioural output; (II) the constancy mechanism condition – the state still represents the external feature despite the variety of proximal stimuli that reach the system's sensory apparatus.

CHAPTER 5. THE CONTENT PROBLEM:

IN DEFENCE OF PRODUCER-BASED TELEOSEMANTICS

The topic of this fifth and final chapter is the content problem: provided that a given state is representational, what determines its content, i.e., what it is about? In virtue of what is the content of a given representational state *C* rather than *C'*? The teleosemantic basic framework claims that the biological function of the representational state determines its content. However, this is just the starting point of the debate on the teleosemantic approach to content. Several functional indeterminacy problems threaten its viability in determining content in terms of the representational state's biological function. As a result, the basic teleosemantic framework may be developed in several conflicting ways. Consumer-based and producer-based teleosemantics are the main teleosemantic approaches to representational content. They deliver different content assignments for the same representational state in paradigmatic indeterminacy cases. My goal in this chapter is to develop a variation of producer-based teleosemantics to determine content and thus to solve these functional indeterminacy problems.

In the first section, I introduce the consumer-based and producer-based teleosemantics and how they deliver distinct representational contents. After that, I introduce the functional

indeterminacy problems and why they are problematic for teleosemantics. In the second section, I assess the first functional indeterminacy problem – the concertina problem – and propose the lower-level thesis to solve it. This thesis is compatible with both consumer-based and producer-based teleosemantics, i.e., both approaches may adopt it to determine content. However, such solution is challenged by a problem that threats to render content relative – the relativity problem. I reject Papineau's consumer-based response to the relativity problem and propose my own producer-based response to it and argue that it solves this problem. I take this to constitute an argument for producer-based teleosemantics. In the third section, I develop an argument for producer-based teleosemantics based on the plausibility of malfunctioning statuses of detection systems and I defend this approach from the objection that it fails to keep the room open for enough misrepresentation cases. In the fourth section, I propose a solution for the second functional indeterminacy problem – the distality problem – that is especially problematic for the producer-based approach. Finally, in the last section I defend producer-based teleosemantics from one final objection – the source of error objection.

5.1 Producer-based x consumer-based teleosemantics

In the second chapter, I introduced the basic teleosemantic framework according to which the content of the representational state is determined by its biological function. That is, the representation's truth-conditions are derived from its biological success conditions. Since the aetiological conception of function is assumed here, the biological function is the effect for which the trait was selected. However, this basic framework is just the starting point of the teleosemantic approach to content. How to properly develop it is a very contentious matter. Several teleosemanticists have proposed different ways of developing it in order to address the content problem. The various teleological theories may be classified into two different approaches according to how they develop this basic framework: the consumer-based and the

producer-based teleosemantics. However, this is not clear-cut. Rather, there are intermediate cases – teleosemantic theories that stay in the middle between these two approaches (e.g., AGAR, 1993). Let's introduce consumer-based and producer-based teleosemantics by showing how they deliver distinct contents in the classic example of the frog's representation of the fly.

A normal frog will snap its tongue at anything suitably small-dark-moving thing regardless of whether it is a fly or not. In the frog's natural habitat, the small-dark-moving things are mostly nutritious flies and so normally the frog will catch a nutritious fly. But the frog cannot discriminate between flies and other small-dark-moving things, it simply responds to anything small-dark-moving which passes its retina by darting its tongue out.[103] So, what is being represented by the frog? What is the function of the frog's visual system – to token the representational state when there is a small-dark-moving thing, fly or food?[104] It is not clear what is actually represented by the frog. It is indeterminate whether it is properly representing or misrepresenting when it detects a SDMT that is not a nutritious fly and vice versa. How to give an account of this case? It depends on how the basic teleosemantic framework is developed.

The first formulation of this framework is the *consumer-based teleosemantics* that was introduced in the second chapter.[105] It claims that content is determined by the biological function of the system that consumes the representational state. Consider the representational state R that triggers the behavioural output B. The content of R is the external condition C that should be the case for B to succeed in achieving the effect that the system which produced B was selected to achieve. That is, it is the biological function of the system that uses the

[103] For the original source of the frog's example, cf. LEETVIN et al., 1959. The appeal to the frog's example to illustrate teleosemantics was made famous by Jerry Fodor, cf. FODOR, 1990. Sometimes the preferred example is the prey-catching toad – frogs and toads have very similar visual systems. For a very empirically detailed presentation of the toad's system, cf. NEANDER, 2006.

[104] "Small-dark-moving things" is just an umbrella term to cover all the properties of the fly to which the frog's visual system is sensitive. From now on, I will use the abbreviation "SDMT".

[105] Cf. PAPINEAU, 1998, 2003, 2016; MILLIKAN, 1989b; 2004; 2007; PRICE, 1998, 2001.

representational state that determines content. The represented external condition is the one required for the behavioural output produced by the consumer system to perform its function. So, in the case of the frog's representational state, its content is *frog's food* or *nutrients*. The reason is that the external condition required for the consumer system (the motor and digestive systems that catch and digest the fly) to have its adaptive effect (the digestion of the nutrients) is that the represented object is nutritive. If the represented object is not nutritious, then this is a misrepresentation since the state is representing the presence of a nutritive object. After all, the caught and digested object should be nutritious for there to be increase of fitness. In sum, the state represents the adaptive properties of the represented object.

However, consumer-based teleosemantics is rejected by several teleosemanticists who develop the basic teleosemantic framework into the opposite direction. They claim that content should be determined not via the selected effect of the consumer system, but via the selected effect of the producer system. That is, it is the biological function of the producer system that determines content. This is the *producer-based teleosemantics* (also called "informational teleosemantics").[106] The selected effect that constitutes the producer's function is the detection or tracking of a given external condition. Such detection consists in the tokening of the representational state whenever this external condition obtains.[107] It is this selected effect that determines the external condition that constitutes content – the one that the producer system was selected to discriminate. The function of the producer system is to be causally sensitive to this external condition. That is, the producer system was selected to have the capacity to discriminate a certain external condition and it is this selected effect that determines content. The producer's function is the selected effect of tokening the representational state in response

[106] Cf. DRETSKE, 1986, 1988, 1995; NEANDER, 1995, 2013, 2017; JACOB, 1997; SCHULTE, 2012.

[107] That Is the reason that the detection of an external condition is a genuine effect of the producer system. For the system to detect C is just for it to token the representation whenever C obtains. But it is perspicuous that to do something in response to another thing is a genuine effect of the system. If the system was selected to have such response, it has what Neander calls a "response function" (NEANDER, 2017, p. 125-47).

to the presence of the external condition that the producer was selected to discriminate. The state represents the properties of the object that the producer system has the function to discriminate.[108]

Let's apply the producer-based approach to the frog's example. The selected effect of the frog's visual system is its capacity of discriminating certain external condition – the presence of a SDMT. Its function is to token the representation whenever these visual properties are present. The content of the frog's representational state is *SDMT* because the adaptive effect of the visual system is to token the representational state whenever there is the presence of SMDTs. The reason is that the visual system that produces the representation was not selected to discriminate nutritive properties. In fact, it has no capacity of discriminating nutritive properties at all. Rather, it was designed to discriminate SDMTs, these are the properties to which it is sensitive. It was via the detection of SDMTs that the motor and digestive systems were able to catch and digest nutritive flies because in the frog's historical environment there was a strong correlation between SMDTs and nutritive flies. There is malfunction when the producer system fails to fulfil the effect for which it was selected – the discriminatory capacity of tokening the representation whenever certain external condition is present. There are two situations in which it may happen – when the system tokens the representation but the external condition does not obtain (misrepresentation) or when it fails to token the representation even though the condition obtains (ignorance). In both cases, the producer system is malfunctioning.[109]

There is a very contentious debate between producer-based and consumer-based teleosemantics. Their fundamental distinction arises when considering minority cases in which

[108] Producer-based teleosemantics is also called "informational teleosemantics" because its core thesis may be characterized in terms of the notion of information. "The fundamental idea is that a system, S, represents a property, F, if and only if S has the function of indicating (providing information about) the F of a certain domain of objects." (DRETSKE, 1995, p. 2).

[109] Does producer-based teleosemantics really succeed in keeping the door open for malfunctioning and so for misrepresentation? This is a common objection that I assess in the third section.

these rival approaches deliver different truth values to the relevant representational state. They usually deliver the same truth value to the representation, but there are minority cases in which the representational state is true according to one approach and false according to the other or vice versa. These are the situations in which the external condition that is represented according to the consumer-based approach is not satisfied but the external condition that is represented according to the producer-based approach is satisfied or vice versa.

Suppose that some scientist fools the frog by placing a small-black-moving pellet in front of it and as a result the frog catches and digests the pellet. The problem that arises is whether the frog misrepresents the pellet or not. According to the consumer-based teleosemantics, the state represents the nutritive properties of the object and given that the small-black-moving pellet is not nutritive, it follows that this is a misrepresentation. By contrast, the producer-based teleosemantics claims that the state represents the sensitive properties of the object – blackness, smallness and movingness – and so it follows that the state accurately represents reality since the pellet is in fact a SDMT. This is a case in which the relevant object is a SDMT but is not nutritive and as a result the representation is true according to producer-based teleosemantics but false according to consumer-based teleosemantics. The same thing happens in case the scientist places a nutritive pellet that is not a SDMT in front of the frog. As a result, the representation is false according to producer-based teleosemantics (e.g., in virtue of an internal defect, the visual system tokens the representation even though there is no SDMT around), but true according to consumer-based teleosemantics. So, how to decide between these two conflicting and rival approaches? How to assess the debate between them to decide which one is the right teleosemantic theory?

The minority cases in which producer-based and consumer-based teleosemantics deliver distinct truth-values to the representational state illustrate quite well the fundamental disagreement between them. But what is the source of disagreement? Some philosophers claim

that the distinction between producer-based and consumer-based teleosemantics arises from how they respond to functional indeterminacy problems – different responses give rise to producer-based and consumer-based teleosemantics. So, Neander claims that these approaches arise from different responses to the concertina problem. She establishes a functional indeterminacy case which threats the viability of the teleosemantic core thesis that biological function determines representational content (NEANDER, 1995, pp. 124-30).

However, here I will develop a different assessment of the debate between producer-based and consumer-based teleosemantics. They arise not from responses to functional indeterminacy cases, but from the establishment of different criteria to specify the biological function that determines content. Producer-based teleosemantics proposes the criterion that it is the producer system's function that determines representational content, while consumer-based teleosemantics proposes the criterion that it is the consumer system's function that determines content. These distinct producer-based and consumer-based criteria are the source of the fundamental disagreement between these approaches. Producer-based and consumer-based are both threatened by the concertina problem and other functional indeterminacy problems, however their fundamental divergence is not originated from different responses to these problems. But before showing why this the right approach to assess the debate between producer-based and consumer-based teleosemantics, let me introduce what functional indeterminacy problems are and why they are fundamentally problematic for teleosemantics.

Functional indeterminacy cases are the ones in which it seems that there are equally strong reasons to describe the function of a given trait in different and conflicting ways. As a result, there is an indeterminacy in the function of the trait and hence in the malfunctioning or proper functioning status of the trait. That is, it is indeterminate if the relevant trait is proper functioning or not and so there are conflicting judgements on its proper functioning status. Since the teleosemantic basic framework claims that biological function determines content,

functional indeterminacy entails content indeterminacy. If truth-conditions are derived from biological success conditions, then indeterminacy of biological success conditions entails indeterminacy of truth-conditions. The function of a given trait is the effect for which it was selected, and so functional indeterminacy consists in the indeterminacy of the trait's selected effect. That is, it is indeterminate which effect was the selected one. Functional indeterminacy is just selected effect indeterminacy. These are cases in which among the effects of a given trait, there are reasons to determine that a given effect was the selected one while there are other equally strong reasons to determine that another effect was the selected one.

Functional indeterminacy cases are problematic for teleosemantics because there is content indeterminacy when it is indeterminate which effect of the representational state is the selected effect that determines content. Suppose that a given representational state has two effects, E and E*, and it is indeterminate which one is the selected effect that determines content because there are strong reasons to hold that E or E* is the selected effect. It is thereby indeterminate what its truth conditions are. That is, it is indeterminate which situations the representation is true and which situations it is false. There are specific situations in which E is achieved but not E*, so it is indeterminate in these situations the representation's truth value. The representation is true whether E is the effect that determines its content, but it is false whether E* is the effect that determines its content. Hence, the problem that functional indeterminacy cases gives rise to teleosemantics is precisely how to determine that this effect but not that effect was selected and so that the former but not the latter effect determines content.[110]

Neander claims that producer-based and consumer-based teleosemantics originate from different responses to functional indeterminacy cases. They arise from the establishment of two

[110] The problem of functional indeterminacy for teleosemantics was developed by Fred Dretske and Jerry Fodor, cf. DRETKSE 1986; FODOR, 1990.

criteria that specify different selected effects of the representational state and as a result deliver different functions and contents. Producer-based teleosemantics claims that it is the function of the producer system that determines the function of the representation – the latter is derived from the producer's function. Consumer-based teleosemantics claims that it is the function of the consumer system that determines the function of the representation – the latter is derived from the consumer's function. According to Neander, the establishment of these criteria in order to solve functional indeterminacy problems give rise to producer-based and consumer-based teleosemantics. However, this is not the case.

The starting point of the disagreement between producer-based and consumer-based teleosemantics is not the disagreement on what is the selected effect of the representational state in order to determine its function and hence its content. Rather, the disagreement arises because they focus on the function of different systems to determine content. According to the producer-based approach, it is the function of the producer system that determines content; according to the consumer-based one, it is the function of the consumer system that determines content. That is the fundamental disagreement between these two approaches and the reason that Neander's assessment of the debate between them is inadequate. It is not that these approaches deliver different content in virtue of their disagreement on what is the function of the representation in functional indeterminacy cases. The above fundamental disagreement would still stand even if there were no functional indeterminacy cases on the function of the representation, of the consumer or producer system, or of any other biological trait. Producer-based and consumer-based teleosemantics disagree from the very start, even before functional indeterminacy cases arise to put into question their viabilities.

Note that I am not denying that there is a disagreement between producer-based and consumer-based teleosemantics on what is the function of the representational state and hence on the delivered contents. Rather, I am arguing that such disagreement is just a by-product of

the above fundamental disagreement. Both teleosemantic approaches assume that it is the function of the representational state that determines content and that the function of representational state is derived from some system of the organism. However, they disagree on which is the relevant system. Producer-based teleosemantics claims that the function of the representational state is derived from the producer system's function, while consumer-based teleosemantics claims that it is derived from the producer system's function. So, the disagreement between consumer-based and producer-based teleosemantics on the function of the representational state is merely derived from their fundamental disagreement on the relevant system whose function determines the representational state's function and content.

But if the fundamental disagreement between producer-based and consumer-based teleosemantics is not originated from different responses to functional indeterminacy problems, what is the role of these problems on the debate between these teleosemantics approaches? Functional indeterminacy threats to render indeterminate the content delivered by both producer-based and consumer-based teleosemantics by threating to render indeterminate the functions of the consumer and producer systems. The indeterminacy of the selected effect of the producer or consumer systems entails content indeterminacy. My strategy in respect of this debate will consist in taking as starting point the functional indeterminacy problems that threatens both producer-based and consumer-based teleosemantics in order to assess whether they are doomed or not by these problems. I will try to show that producer-based teleosemantics has the resources to develop solutions to these problems which will constitute arguments in favour of producer-based teleosemantics in detriment of consumer-based teleosemantics. The strategy will be to use the functional indeterminacy problems that threaten both teleosemantic approaches as a common ground in order to show that producer-based teleosemantics should be favoured in light of its success in solving these problems.

But what are the concrete cases of functional indeterminacy that threatens producer-

based and consumer-based teleosemantics by threating to render content indeterminate? That is, the cases in which there are equally strong reasons to claim that this or that effect determines the function of the producer or consumer systems? As it happens, there are several sources of functional indeterminacy that give rise to different functional indeterminacy problems. "Functional indeterminacy problem" is just an umbrella term to cover all these specific problems. However, it is not my goal here to assess all of them. In this chapter, I assess two functional indeterminacy problems. In the next section, I assess the concertina problem and in the fourth section, I assess the distality problem. I argue that producer-based teleosemantics solves these problems and this constitute a reason in favour of this approach.[111] I also develop a general defence of producer-based teleosemantics in the third section. Finally, in the last section I assess and reject a final objection to this approach – the source of error objection.

5.2 Functional indeterminacy (I): the concertina problem

Antelopes are mammals that live at lower ground. Suppose that a trait in this population altered the structure of the haemoglobin which caused higher oxygen uptake which finally allowed antelopes to survive at higher ground. That was adaptive because the antelopes were forced to move to higher ground. Now suppose that as a result this trait was selected by natural selection. The problem that arises is what was the effect of the trait in virtue of which it was selected? The altered haemoglobin's structure? The higher oxygen's uptake? The survival at higher grounds? All effects, actually. For all of them were done and were adaptive. Indeed, these effects were not achieved independently, but it was by achieving one effect that the others were achieved. The antelope's trait (IV) contributed to survive and reproduction; by → (III)

[111] Jerry Fodor is famous for developing a functional indeterminacy problem for teleosemantics according to which biological function cannot determine representational content because natural selection is extensional while content is intensional (FODOR, 1990). I will not assess this problem here, but I think that Elliott Sobber's solution based on the distinction between selection-for and selection-of a trait solves it. Cf. SOBER, 1984, 2008, 2010.

allowing the antelope to survive at higher ground; by → (II) increasing oxygen uptake; by → (I) altering the hemoglobin's structure. As this sequence shows, the trait does one thing by doing another.

The lower description (I) describes what the antelope's trait effectively does, and the higher levels (II) and (III) explain why doing that was adaptive. The trait alters the structure of the haemoglobin and it is adaptive because it increases the antelope's oxygen uptake and hence allows the antelope to survive at higher ground. Notice that as we move up the diagram from the most fundamental level to higher levels, we are moving to descriptions of functions of larger and larger systems. To alter the haemoglobin's structure is something that the trait does in a more or less independent way, but to increase oxygen uptake and to move the antelope to higher ground demand the help of other traits – the respiratory and the motor systems. But in which level of the sequence lies the right description of the biological function of the antelope's trait? That is, which one is the appropriate functional assignment? That is the concertina problem (DRETSKE, 1986; NEANDER, 1995). This is a general problem for the aetiological conception of biological function – there is a concertina of selected effects that a given system should have and it is indeterminate which selected effect constitutes its biological function. So, the problem is to specify which effect among the concertina of effects of the trait constitutes its biological function.[112]

It is easy to show why the concertina problem threatens the viability of teleosemantics. Let's show it by appealing to the frog's example again. The frog's representational state triggers the frog's behaviour of snapping its tongue, catching the fly and finally digesting it. So, what is the function of the frog's visual system that produces the representational state?

[112] Dretske was presumably the first to formulate the concertina problem (DRETSKE, 1986). However, Neander was the first to explicitly formulates the concertina problem as arising from a distinct source of functional indeterminacy from other sources and so to show that it is a distinctive functional indeterminacy problem (NEANDER, 1995). This problem is also called "the problem of the complex causal role" (NEANDER, 2012).

Once again, there are several adaptive effects here. The frog's visual system (IV) contributed to gene replication; by → (III) helping to feed the frog; by → (II) helping the frog to catch flies; by → (I) detecting SDMTs (flies? food?). So, has the frog's visual system the function of detecting SDMTs, flies or food? There are equally strong reasons to describe the function of the visual system as to detect SDMTs, frogs or food. That is, the function seems to be adequately described in these different and conflicting ways. All effects were achieved and were adaptive, it was by doing one thing that the others were done. In the same vein, the consumer system has a concertina of effects and it is indeterminate which one constitutes its function – to catch and digest SDMTs, flies or food? The result is that for both producer-based and consumer-based teleosemantics it is indeterminate what the content of the representational state is – *SDMT, fly* or *food*? Solving the concertina problem is a demanding task for the teleosemanticist – i.e., determining the producer and consumer systems' functions and hence determining the content of the representational state.

I think that the first step to solve the concertina problem is to provide a functional analysis of the organism to which the trait pertains. Robert Cummins, inspired by the conceptual analysis of an organism by biologists, proposes the following functional analysis (CUMMINS, 1975). In physiology, the frog's organism is decomposed into the major physiological systems (circulatory, digestive, nervous, etc.) and the contribution made by each system is specified. Then, each system is decomposed into its parts (e.g., the digestive system into the mouth, stomach, etc.) and it is specified the contribution made by each part to the system of which it is a component. These parts, by turn, are then decomposed into their parts (the mouth in the tongue, saliva glands, etc.) and again the contribution made by each part is specified. This decomposition process will continue down to the level of individual cells and their sub-cellular components, with the specification of the causal contribution of each decomposed part.

The functional analysis illustrates the fact that the effects of a trait depends not on the trait alone, but also on other traits of the physiological system as well as on other physiological systems. For instance, the digestion of the fly is a joint effect of the frog's motor and digestive systems. The contribution of the motor system to the whole consumer system is to catch the fly, while the contribution of the digestive system is to digest the fly. However, the functional analysis alone doesn't specify the effect which constitutes the function of the trait among its concertina of effects.

The second step to solve the concertina problem consists in the proposal of a criterion to specify in which level of the analysis the trait's function should be determined. Here I will defend *the lower-level thesis* – the function of a biological trait is its specific or direct effect. I think that Neander proposed a good strategy to support this thesis by switching the focus from proper functioning to malfunctioning cases in order to show that the function of a trait is its most specific or direct effect (NEANDER, 1995). Let's focus on the plausibility of malfunctioning statements: what is the right criterion to determine when a trait is malfunctioning or not?

Consider the antelope's trait again. It contributes to fitness by allowing the antelope to survive at higher grounds by increasing oxygen uptake and finally by altering the haemoglobin's structure. They are all effects of the trait and their fulfilments imply the proper functioning status of the trait. But which one's absence is responsible for its malfunctioning? Suppose that the motor system fails to move the antelope to the higher ground and as result there is no contribution to gene replication in virtue of an internal defect of the motor system. In this case, it is wrong to claim that the antelope's trait is malfunctioning because the effect of moving the antelope to the higher ground was not fulfilled, after all this failure was caused by an internal defect of the motor system. The trait has no responsibility for this failure at all.

This case illustrates the fact that most of the effects of a trait depend not only on the

proper functioning of the trait, but also on the proper functioning of other traits. The movement of the antelope to the higher ground is a joint result of the relevant trait and the motor and respiratory systems. The function of a trait is the effect for which it is directly responsible, i.e., the effect whose fulfilment depends on the trait alone. If the trait fails to have this specific effect, then it is malfunctioning. So, the criterion to determine the function of a trait lies on the effect for which it is directly responsible. The trait is malfunctioning if and only if it fails to have its direct effect and the trait is properly functioning if and only if it succeeds in having its direct effect.

The third step to solve the concertina problem consists in the specification of the trait's specific or direct effect that constitutes its biological function. But how to specify it? The trait's direct effect is the effect that appears in the lowest level of the functional analysis in which the trait appears as an unanalysed component (NEANDER, 1995, pp. 129-30). The nature of the unanalysed component is explained in this way. Consider the functional analysis of an organism *O* in which in the first stage *O* is decomposed into the parts *Oa*, *Ob* ... *On*; in the second stage *Oa* is decomposed into the parts *Oa1*, *Oa2* ... *Oan* and that in the third stage O*a1* is decomposed into the parts *Oa1a*, *Oa1b* ... *Oa1n*. Note that *Oa1* appears as an unanalysed component in the second stage but not in the third stage. In the third stage, *Oa1* is analysed into the parts *Oa1a*, *Oa1b*... while in the second stage *Oa1* is one of the unanalysed components (along with *Oa2*, *Oa3*...) that constitute the parts of the functional analysis of *Oa*.

The function of the frog's visual system is its direct effect, namely, the effect that appears at the lowest level in which the visual system as a whole appears as an unanalysed component in the functional analysis. That level is the functional analysis of the frog's sensory system, a physiological system which is decomposed into the visual system, olfaction system, auditory system, etc. As showed by the effects sequence, the visual system has a concertina of effects, but some effects are in higher levels than others. In this level of the functional analysis,

the effect of detecting the object (the SDMT or fly or food) is in the lowest level of the effect's diagram, while the effect of catching the fly and digesting it are in higher levels. It is via this detection that the visual system helps the frog to catch the fly and digest it. Thus, the function of the frog's visual system is to detect the object. That is the specific effect of the system, the only effect for which it is directly responsible. In the same vein, the function of the frog's motor system is to catch the object, not digest it, because this is its direct effect.

In sum, to determine the effect that constitutes the function of a given trait among a concertina of effects, three steps are required. The first is to provide a functional analysis of the organism in which it is a component. The second step – the lower-level thesis – is to identify the trait's function with the trait's direct effect. The third and final step is to specify the trait's direct effect as the trait's effect that appears at the lowest level of the functional analysis in which the trait appears as an unanalysed component. The function of the trait is this direct effect.

So far, so good. But at this point arises this problem: is it the function of the frog's visual system to detect SDMTs, flies or food? That is, among the visual system's concertina of effects, which one is its direct effect? Neander argues that it is to detect SDMTs because "it is by detecting small dark moving things that the frog detects frog-food and flies" (NEANDER, 1995, p. 130). However, this argument is problematic. The visual system does not detect one thing by detecting another thing. This is not the case because the system just detects one thing and what is at issue is precisely how to appropriately describe what it is its function to detect. That is, the question is to specify which one is the right detection assignment – SDMT, food or fly? These are not successive detection effects, but alternative descriptions of the same detection effect. That is the fundamental reason that Neander's argument is flawed.

It is certainly the case that the frog catches and digests nutritive flies via the detection of SDMTs simply because in the frog's historic environment there is a strong correlation

between SDMTs and nutritive flies – the great majority of SDMTs are nutritive flies. However, it is not the case that the visual system detects flies or food via the detection of SDMTs. It either detects one thing or detects another. So, it is required a further reason for the conclusion that the direct effect of the visual system is to detect SDMTs – not flies or food. The lower-level thesis does not rule out that the direct effect of the frog's visual system is to detect flies or food. It leaves open whether the visual system's function is to token the representational state in response to the presence of SDMTs, flies or food. So, how to determine the detection function of the visual system?

In this chapter, I develop a variation of the producer-based teleosemantics according to which the function of the frog's visual system is to detect SDMTs, the function of the motor system is to catch SDMTs and the content of the representational state is *SDMT*. I develop a defence of the producer-based approach that appeals to the lower-level thesis but since this thesis does not imply this approach – it is also compatible with consumer-based teleosemantics – further arguments are required in favour of producer-based teleosemantics. The first argument is developed in what remains of this section. The lower-level thesis is compatible with both producer-based and consumer-based teleosemantics, but there is a problem that threatens the viability of the adoption of this thesis by both approaches – the relativity problem. It threatens to render content relative. I then assess and reject a consumer-based solution for this problem developed by David Papineau and after that I propose my own producer-based solution for the relativity problem. I take the solution of this problem to constitute an argument for producer-based teleosemantics to the detriment of consumer-based teleosemantics. Let's start by showing how producer-based and consumer-based teleosemantics are compatible with the lower-level thesis – both may appeal to it in order to determine function and representational content.

The relativity problem.

Producer-based teleosemantics gives priority to the lowest level of the effects sequence and claims that the function of the producer system is the detection of a given external feature – the one that the producer system was selected to be sensitive to. In the case of the frog, the function of the visual system is to detect SDMTs because it was designed to have the capacity of discriminating SDMTs. This is the stimulus that triggers the frog's catching behaviour. It has no capacity for discriminating nutritive properties and hence it is not its function to detect food. According to producer-based teleosemantics, it is the producer's function to discriminate certain external feature that determines content. Applying the lower-level thesis to this producer-based approach, it follows that the visual system's function is to detect SDMTs and hence that the content of the frog's representational state is *SDMT*, not *fly* or *food*.

By contrast, according to consumer-based teleosemantics it is the function of the consumer system that determines the content of the representational state – the content is the external condition that should be the case for the consumer system to perform its function. Papineau has proposed a variation of the consumer-based teleosemantics that assumes the lower-level thesis in order to determine the function of the consumer system (PAPINEAU, 2003, 2016). His idea is to apply the lower-level thesis to determine the function of the consumer system and hence to fix content. The consumer system has a concertina of selected effects: to catch the fly, to swallow it to the stomach, to digest it, etc. However, only the first one is its direct effect, so the consumer system has the function of catching the fly. Notice that according to consumer-based teleosemantics, it cannot be the function of the consumer system to catch SDMTs because what contributes to fitness is the catching of flies, not SDMTs. The consumer system should catch that object that is required for it to contribute to fitness – flies, not SDMTs. So, the representational content is that external condition that should be the case for the consumer system to achieve this direct effect, namely, the presence of the fly. Hence,

the content of the representational state is *fly*, not *SDMT*.

Unfortunately, there is a problem that threats the application of the lower-level thesis by both consumer-based and producer-based teleosemantics – the relativity problem. The application of the lower-level thesis to the case of the frog unwarrantedly assumes that the frog's representational state is part of the visual system. But it is plainly arbitrary to assume it without some previous justification (PAPINEAU, 2003). Why not assume that the frog's representation is part of the digestive system? After all, it is the digestive system that digests the fly. Why not assume that the frog's representation is part of the circulatory system? After all, it is the circulatory system which circulates the nutrients of the fly. It seems that there is no principled way of maintaining that the frog's representation is part of a given physiological system rather than another. The problem for the lower-level thesis is that each physiological system has a function of its own, and so the function of the frog's representation is relative to the system that the representation is being viewed as part of. So, the relativist conclusion that the function of the representation is indeterminate once more.

In light of the relativity problem, it could be argued that there is no full-blown functional indeterminacy here. Rather, there is mere relativity of functional assignment. It is not possible to assign function to a mental representation in absolute terms, regardless of which physiological system the representation is being considered as a component. The function of the representation is relative to which system it is being viewed as a component and there is no fact of the matter of the function of the representation in absolute terms. So, the function of the frog's representation is relative to different physiological systems: to detect SDMT (relatively to the visual system); to detect flies (relatively to the motor system); to detect stomach filler (relatively to the stomach system); to detect food (relatively to the digestive system); etc.[113] The problem with this position is that the content of the representational state cannot be relative

[113] Papineau once embraced this relativist position, cf. PAPINEAU, 2003.

to the physiological system because a relative content cannot play any role in the explanation of the behaviour of the organism. So, there is no explanatory relevance of attributing relative contents to mental representations. Content cannot be relative for the representational state to play a role in the explanation of behaviour.

So, the establishment of a criterion to determine of which physiological system the frog's representational state is a component is required to avoid the relativist conclusion. The producer system is the physiological system which produces the representation – the frog's visual system. The consumer systems are the physiological systems which use or consume the representation, i.e., the systems which historically used or consumed the representation in order to fulfill their own functions. In the case of the frog, the motor system is a consumer system inasmuch as it uses the representation to catch the fly; the digestive system is a consumer system inasmuch as it uses the representation of digest the fly, etc. Of which system is the representational state a component?

Papineau's solution

Papineau has recently rejected his former relativist position and proposed that the frog's state is part of the prey-catching system and so that its content is *fly* (PAPINEAU, 2016, p. 106-7). Let's assess his argument in detail. It claims that the representational state is properly seen as a component of the prey-catching system – the visuomotor system that governs head-turning and tongue-snapping – not of any other system. Consider the larger prey-stomaching system, constituted by the prey-catching and the prey-swallowing systems. When analysed, there is no requirement of bringing in the representational state. That is the case because the prey-stomaching system fulfils its function as long as its constituting prey-catching and prey-swallowing systems fulfil their functions – no matter how they do that. So, there is no requirement of bringing in the representational state when analysing the prey-stomaching

system. It is only when the prey-catching system is analysed that the representational state appears – this system fulfils its function as long as it head-turns and snaps the tongue in the direction in which there is in fact the fly. So, the proper functioning of the prey-catching system requires that the representational state tracks the external condition in which there is in fact a fly. The representational state lies between the visual system that produced it and the motor system that governs the head-turning and tongue-snapping. So, the conclusion is that the representation is specifically a component in the prey-catching system, its function is to detect the presence of the fly in a certain direction.

What is problematic with this argument? It fails to show that the representational state is a component in the prey-catching system; that this is the system where it appears when properly analysed. The problem is that the argument fails to show that the representational state is part of the prey-catching system in opposition with the prey-stomaching or the prey-digesting systems. The reason appealed to by the argument to show that the representation is part of the prey-catching system also implies that it is part of the prey-stomaching and prey-digesting systems. My point is not that the representation is not a component of any physiological system. Evidently, it is. Rather, my point is that the argument fails to show that the representation appears between the visual system and the motor system but that it does not appear between the visual system and the prey-stomaching or between the visual system and the prey-digestive system. Let me show this in detail.

It is obvious that in the functional analysis the representation appears as soon as the visual system appears, precisely because the latter produces the former. So, the representation appears between the visual system (the producer system) and some consumer system. However, the above argument fails to specify which one is this consumer system. Is it the motor system and as a result the representation is a component in the prey-catching system? Or is it the motor-swallowing system and as a result the representation is a component in the larger prey-

stomaching system? Or is the relevant consumer system the one constituted by the motor system and the whole digestive system and thus the representation is a component in the prey-digesting system? They are all systems that consume the representation.

Papineau's argument claims that the representation appears only when the prey-catching system is analysed because it fulfils its function as long as its components fulfil their functions and such fulfilment requires that the representation tracks the presence of a fly that will result in the motor system catching the fly. But this is also the case with the prey-stomaching and prey-digestive systems. The motor-swallowing system and the motor-digestive system also consume the representation. If that is the case, why does only the fulfilment of the function of the components of the prey-catching system requires the tracking of the external condition and not the fulfilment of the components of the prey-stomaching and prey-digesting systems? Papineau seems to presuppose that this is the case because the functional analysis reveals that the motor system is a component of the prey-stomaching system, while the stomaching system is not a component of the prey-catching system. But this fact does not entail that the representation is a component in the prey-catching system in contrast to the prey-stomaching system. It is equally plausible that the representation stands between the visual system and the motor system as well as between the visual system and the larger motor-swallowing system. In both cases, the system fulfils its function as long as the visual system and the relevant consumer system fulfils their functions. The prey-catching fulfils its function as long as the visual system and the motor system fulfil their functions and the prey-stomaching system fulfils its function as long as the visual system and the motor and swallowing systems fulfil their functions. In all these cases, the system fulfils its function as long as the visual system and the relevant consumer system fulfils their functions. The relevant consumer system is equally conceivable as the motor system or the motor-swallowing system.

Finally, Papineau's argument presupposes that the functional analysis and the lower-

level thesis entails that it is the function of the motor system to catch flies, so let's assume for the sake of the argument that its function is in fact to catch flies. Just like the motor system fulfils its direct function as long as it catches the fly, the motor-swallowing system fulfils its direct function as long as it catches and swallows the fly. The representational state appears in both functional analysis – the one in which the representational state stands between the visual and the motor system as well as in the one in which the representational state stands between the visual and the motor-swallowing system.

It could be replied that the appropriate functional analysis reveals that the representational state stands between the visual system and the motor system (not the motor-swallowing system) because when analysed the motor-swallowing system is decomposed into the motor and swallowing systems. The problem with this response is that it does not entail that the representational state should appear between the visual system and the motor system rather than another consumer system. It could be replied that the appropriate functional analysis reveals that the representational state appears as an unanalysed component at the level in which it stands between the visual system and the motor-swallowing system. It is arbitrary to choose one functional analysis rather than another in the absence of some principled criterion. The conclusion is that the argument fails to determine which one is the consumer system of the representation and so fails to determine which system the representation is a component of – the prey-catching system, or the prey-stomaching system, or the prey-digesting system, etc.

The lesson to be taken is that it is hard to conceive of a non-arbitrary and principled criterion that succeeds in showing that the representational state lies between the visual system and the motor system but not between the visual system and some larger consumer system like the prey-stomaching system. But if there is no available criterion to determine which one is the specific consumer system of the representational state, content cannot be determined by the function of the consumer system as proposed by consumer-based teleosemantics. As a result,

content indeterminacy arises once again. That is, it is hard to adopt the functional analysis and the lower-level thesis to determine the function of the consumer system and then to determine content in terms of the consumer's function precisely because it is hard to determine which one is the relevant consumer system and so which function determines content. It is more promising to apply the functional analysis and the lower-level thesis to producer-based teleosemantics according to which the function of the producer system determines content – in the case of the frog, the visual system – precisely because the producer system is fully specified. In what follows, I develop this strategy.

The producer-consumer solution

There are different candidates for the criterion to determine to which system the representational state is a component. The first one is *the producer criterion*: the representational state is part of whatever physiological system which produced this representation. Since the frog's representation is produced by the visual system (not by the digestive system, or by the circulatory system...), it follows from the producer criterion that the frog's representation is part of the visual system (not of the digestive system, or of the circulatory system...). That is the criterion implicitly assumed by Neander (NEANDER, 1995). The second candidate is *the consumer criterion*: the representational state is part of whatever physiological system that consumes it. Since the frog's representation is not consumed by the visual system, it follows from the consumer criterion that the frog's representation is not part of the visual system, but part of the motor system, digestive system and other consumer systems. Finally, from the conjunction of the producer criterion and the consumer criterion arises the third candidate, *the producer-consumer criterion*: the representational state is part of whatever physiological system which produces or consumes it. Since the frog's representation is produced by the visual system and is consumed by the motor system, the digestive system

and other consumer systems, it follows that the representation is part of the visual system, the motor system, the digestive system, etc. Here I will defend the producer-consumer criterion. My strategy is to develop an argument to show that producer-based teleosemantics has the resources to solve the relativity problem by providing a justification for the producer-consumer criterion. It shows that despite the variety of consumer and producer systems of which the representation is a component, its content is the same relative to all of these systems. I take such a solution to constitute an argument in favour of this producer-based approach.

The lower-level thesis shows that the function of the frog's motor and digestive systems are respectively to catch and digest the relevant object since these are their direct effects. But this thesis leaves open which object it is their functions to catch and digest – fly, food or SDMT? It doesn't matter to the function of the motor system whether the represented object is nutritious or not as long as it catches the detected object – the motor system performs its function when it catches a fly as well when it catches a SDMT which is not a fly. It also doesn't matter to the function of the digestive system whether the caught object is a fly or not as long as it digests the object – the digestive system performs its function when it digests a nutritious fly or something else. In the same vein, it doesn't matter to the function of the circulatory system whether the digested object is nutritious or not as long as it circulates the digested object through the bloodstream – the circulatory system performs its function when circulates a nutritious stuff as well as a non-nutritious one. These are the functions of the first three physiological systems which consume the representation. They are proper functioning provided that they respectively catch, digest and circulate the detected object. It is not their fault or responsibility if the producer system fails to detect whatever it is supposed to detect. Rather, it is a fault of the producer system.

What about the producer system, the visual system? The lower-level thesis shows that its function is the detection, not to help the catching of the fly or the feeding of the frog, because

this is its direct effect. But is its function to detect flies, food or SDMTs? The application of the lower-level thesis to the producer-based teleosemantics shows that the function of the producer system is to detect the external condition that it was selected to discriminate, and the frog's visual system was selected to discriminate SDMTs. On the one hand, it has no capacity to discriminate nutritive things that are not SDMTs, only SDMTs. On the other hand, it doesn't matter to the frog's visual system whether the detected object is a fly or not as long as it is a SDMT – it performs its function when it detects a fly as well as other SDMTs. So, the function of the visual system is to detect SDMTs and the content of the representational state is *SDMT* when the representation is viewed as part of the visual system.

But what precisely is the content of the frog's representation relative to each one of the consumer systems? Let's start with the motor system. An alternative is to claim that if the frog's representation is part of the motor system, the content of the frog's representation is *fly*. But this assignment surely doesn't follow from the lower-level thesis that the function of a trait is its direct effect. The direct effect of the motor system is to catch the represented object, i.e., to catch that object which is represented by the frog's representation as being in a certain position. The motor system is not malfunctioning if it catches the object that the representational state dictates it to catch, no matter whether this object is a fly or not. Rather, it is by having its direct effect of catching the object that the representation dictates it to catch that the motor system usually catches the fly. So, the content of the frog's representation is not *fly* when the representation is viewed as part of the motor system. For the same reason, it is not *food*.

But if the content of the frog's representation relatively to the motor system is neither *fly* nor *food*, what is its content? It is *SDMT*. What happens is that the motor system inherits from the visual system the content of the representation. The function of the motor system is simply to catch the object that the representation dictates it to catch and it doesn't matter to its proper functioning the nature of the object as long as it is caught. Based solely on this fact,

there is no determinate content. But since the object is presented by the visual system to the motor system as a SDMT, the motor system represents the object as a SDMT that should be caught. The visual system produces the representation that there is a SDMT in a certain position and dictates to the motor system that this SDMT should be caught. Then, the frog snaps the tongue in that direction and catches the object. It is through this dictation that the motor system inherits the content of the representation from the visual system.

A primitive representational state like the frog's one is a pushmi-pullyu representation, it states that a certain state of affairs is the case and dictates what the consumer of the representation should do.[114] That is, the pushmi-pullyu representation simultaneously describes the world and dictates what the organism should do in this world, i.e., to trigger a certain behaviour. The descriptive aspect of the frog's representation is the statement that there is a SDMT in a certain position and its imperative aspect is the dictation to the motor system to perform its function towards this SDMT, resulting in its capture when properly functioning. It is through the imperative aspect of the frog's representation that the motor system inherits the representational content from the visual system. Such inheritance cannot be done through the descriptive aspect because if there were no imperative aspect, there would be no catching behaviour by the motor system. Indeed, there would be no behaviour at all.

By the same line of reasoning, the content of the representation is *SDMT* when conceived as part of the digestive system. The function of the digestive system is simply to digest the captured object, but since this object is presented by the visual system as a SDMT, it follows that the digestive system represents the object as a SDMT that should be digested. The digestive system inherits the content of the representation from the visual system and this inheritance occurs through the imperative aspect of the representation, in the sense that the

[114] In primitive representations, the imperative and the descriptive aspect come together in an inseparable way, while in complex representations such as propositional attitudes these two aspects come apart (MILLIKAN, 1995).

representation dictates the digestive system to perform its function towards the SDMT, resulting in the digestion of it. Generalizing this line of reasoning to all consumer systems of the frog's representation, they represent the object as a SDMT because the producer system presents this object to them as a SDMT. All consumer systems inherit the content from the producer system and this inheritance occurs through the imperative aspect of the representation, in the sense that the representation dictates each consumer system to perform its function towards the SDMT. Thus, the content of the frog's representation is *SDMT* relatively to all consumer systems – motor, digestive, circulatory systems, etc.

But of which physiological systems is the representational state a part? It is certainly part of the producer system because that is the system which produced it. It is plausible to claim that a representation is part of the system which produced it, after all the representation wouldn't exist at all in the absence of the producer system or when the producer system is malfunctioning. However, that is not the only system to which it pertains. Since the consumer systems use the representation in order to perform their own functions, the representation is part of these systems too. The effects of the representation reach all consumer systems and so it is part of these systems. They use the representation insofar as they need the represented object to perform their functions and when they do it, the object is presented to them in a certain way. Hence, the consumer-producer criterion is the right criterion to determinate of which physiological systems a given representational state is a component.

The content of the frog's representation is SDMT in all cases, when the representation is viewed as part of the producer system as well as when viewed as part of any consumer system. Since the representation is part of the producer and consumer systems and that the content of the representation is *SDMT* relatively to all of these systems, the conclusion is that the content of the representation is *SDMT* in absolute terms, no matter in relation of which physiological system it is being viewed as a component. Thus, there is no relativity of

functional assignment to the frog's representation and the relativist conclusion is fully avoided.

At this point the following objection to this proposal could be raised. It is highly implausible that a given representational state is part of the whole organism. But this is a consequence of the consumer-producer criterion. That is, if the representational state is part of every physiological system that produces or consumes it, then it follows that the representation is part of the whole organism – not of some physiological systems – precisely because every physiological system can be viewed as either producing (the visual system) or consuming (all others) the representational state. For instance, the frog's respiratory and nervous systems consume the representational state as long as it receives nutrients via the bloodstream. Actually, this line of reasoning ultimately leads to the result that every physiological system consumes the frog's representational state and as a consequence of the producer-consumer criterion, the representation is part of all these systems. How implausible or undesirable is this consequence?

This is indeed a consequence of the consumer-producer criterion. However, this initial implausibility fades away when it is realized that to say that the representation is part of the circulatory, respiratory and further systems is just to say that the total consumer system of the representation includes the respiratory and further systems. In the end, all physiological systems consume the representational state since all systems consume the digested nutrients in one way or another and hence the representational state is also part of all systems. Evidently, we usually just talk about the most immediate systems that use the representational state as the consumer systems – e.g., the motor and digestive systems. But this is just because they are the first systems to effectively consume the representational state. Here it could be objected that it is counter-intuitive to claim that the representation is part of all systems that consume them – it is much more intuitive that it is part just of the most immediate system that consume it, namely, the motor system. However, once again the problem is how to establish a principled and non-arbitrary criterion that implies that the representation is just part of the most immediate

system in contrast with the less immediate ones.

Another objection is that there is no guarantee that the content of the representation is the same along the inheritance process from the producer to the consumer system. How to guarantee that this is the case in order to save the consumer-producer criterion? First of all, it is required to specify the ways through which content could be changed along this inheritance process. The first one is the *partition of content*. For instance, the partition of content relatively to the digestive system results in the representational state not having *SDMT* as its content anymore, but just *small-moving thing*. However, no partition is possible here because as previously noted, the content of simple representations is not partitionable. It is in an indivisible package. So, it is not possible to select some properties which appear on it while excluding others. The second way through which it is possible to change the representational content along the inheritance process is through *the creation of a wholly new content*, not an operation in the inherited content. But in this case, the result would be that in fact a wholly new representational state would be created and so the system responsible for this process would be a producer system. But it was previously assumed that the route of the inheritance process is from a producer to a consumer system. So, neither process is a viable way of changing the content. This objection is flawed.

I think that the lesson to be drawn from the relativity problem is that the frog's representation is part of all physiological systems which consumed or produced it, not part solely of the producer system or solely of the consumer systems. A major threat for this position is the risk that the content of the representation changes from system to system – relatively to a given system it has a given content, while relatively to another system it has another content. In this case, the content would be relative again. But the above argument shows that the content of the frog's representation is the same for all physiological systems of which it is part, namely, *SDMT*. So, this threat is inert.

It is not a surprise at all that the content of the representation is the same relatively to the producer and consumer systems when you look for this conclusion in light of the fact that there is a cooperation between the producer system of a representation and its consumer system. The producer and the consumer systems were designed by evolutionary selection to cooperate with each other in such a way that what the producer system does in the end will help the consumer systems and vice versa. The fulfilment of the producer's function helps the consumer systems to perform their functions and vice versa. The case of the frog illustrates this cooperation quite well. In the majority of cases, when the visual system performs its function of detecting SDMTs, it will detect a fly and so the result will be the capture of the fly by the motor system, the digestion of the fly by the digestive system, etc. That is, the fulfilment of the visual system's function helps the fulfilment of the functions of the motor system, digestive system and of the other consumers. On the other hand, when the consumers perform their functions, the nutrients are transported to the visual system which helps it to detect SDMTs in future occasions. The fulfilment of the functions of the consumer systems helps the visual system to perform its function. So, there is a strong cooperation between the producer and consumer systems.

But by saying that in light of the consumer-producer cooperation it is not a surprise that the content of the frog's representation is the same relatively to the consumer and producer systems, I am not providing an argument for this thesis. I think that this cooperation provides an indication of it, but not an argument at all. This is an indication because it is *prima facie* intuitive that in order to have a cooperation between the producers and consumers of the representation, the representation must have the same content relatively to these systems. Nonetheless, someone could argue that this is not a necessary condition for such cooperation. The producer-consumer cooperation doesn't presuppose that the frog's representation has the same content relatively to the producer and consumer systems in cases where the different

contents are co-extensional (e.g., nutritious flies and SDMTs were historically co-extensional in the frog's natural environment). That is the reason that an argument based on the producer-consumer cooperation fails to show that the representation has the same content relative to the consumer and producer systems.

A final objection to my proposal is that it is ad hoc, stating that I have maintained that the content of the frog's representation is the same relative to the consumer and producer systems just to avoid the relativist conclusion. However, this objection is flawed not just because of the fact that there is a cooperation of consumer and producer systems which shows that it is plausible to think in these terms, but mainly because there is an independent argument for the consumer-producer criterion: the content of the representational state is the same relatively to the consumer and producer systems because the consumer systems inherit the content of the representation from the producer system through the imperative aspect of the representation. If this argument is cogent, as I have tried to show, then this solution to the relativity problem is not ad hoc or arbitrary.

The assumption of the lower-level thesis to determine content in light of the concertina problem is compatible with both consumer-based and producer-based teleosemantics. However, the relativity problem is genuinely problematic for the consumer-based approach because it fails to provide a solution for this problem and, as previously argued, it is hard to conceive such solution. By contrast, producer-based teleosemantics has the resources to solve the relativity problem by providing a justification for the producer-consumer criterion that blocks the relativity threat. The consumer systems defer to the producer system and since there is only one producer system, there is no threat of relativity of content. But several systems consume the representation, and this is problematic for the consumer-based approach – which determines content in terms of the consumer system's function – because that gives rise to the relativity of content threat. So, my conclusion is that the solution of the relativity problem

constitutes an argument for the producer-based teleosemantics and a challenge for the consumer-based teleosemantics. This is my first reason for the adoption of the producer-based teleosemantics. In the next sections, I develop a general defence of this approach and defend it from several objections.

5.3 A defence of producer-based teleosemantics

The lower-level thesis is compatible with both producer-based and consumer-based teleosemantics. It may be assumed by both teleosemantic approaches to determine function, but it does not imply either. Rather, it implies that the producer system's direct effect is the detection of some external condition, but it leaves open which condition is that. So, which of the competing rival descriptions is the appropriate one – the detection of SDMTS, flies or food? How should it be decided?

Consider the following line of reasoning. The direct effect of the frog's visual system is to detect SDMTs but not flies or food because the detection of nutritive flies is dependent on the external correlation between SDMTs and nutritive flies, while the detection of SDMTs depends only on the discriminatory capacities of the visual system. That is, the direct effect of the visual system is the detection of SDMTs, not the detection of nutritive flies. The reason is that the latter depends on the correlation between SDMTs and nutritive flies, while the former is not dependent on this correlation but only on the visual system's discriminatory capacity – it can discriminate only SDTMs. Assuming the lower-level thesis, the conclusion is that the function of the producer system is to detect the properties that it has the capacity of discriminating – SDMTs. How plausible is this argument?

I think that it suggests and constitutes a good evidence for producer-based teleosemantics in contrast with consumer-based teleosemantics. However, it strikes me that a stronger argument is required for producer-based teleosemantics, one that goes beyond the

level of mere evidence or suggestion. That is my goal in this section. Here I propose an argument for producer-based teleosemantics based on the assignment of malfunction statuses to detection systems. While developing it, I assess the objection that producer-based teleosemantics fails to keep room open for enough misrepresentation cases. In the next two sections, I assess two threats to the viability of teleosemantics to give an account of representational content. In the next section, I assess another functional indeterminacy problem that is especially problematic for producer-based teleosemantics – the distality problem. In the final section, I assess the last objection – the source of error objection. The conclusion is that both threats fail to show that the producer-based approach is not viable.

Producer-based teleosemantics claims that the producer system is malfunctioning when it fails to detect the external condition that it was selected to discriminate. A common objection is that such approach fails to keep the room open for misrepresentation cases. That is, producer-based teleosemantics rules out the possibility of the producer system producing false representations – cases in which the representational state is tokened but the relevant external condition does not obtain. This is certainly not the case, but it is important to assess this objection because it gives rise to a problem that threatens the viability of producer-based teleosemantics – does it keep the room open for enough misrepresentation? After all, if it is too restrictive in allowing misrepresentation, then it is not a viable account of representational content.

The function of the producer system is to token the representational state in response to the presence of some external condition of which the producer system was designed to discriminate – the content is that external condition. That is the core thesis of producer-based teleosemantics. The possibility of malfunction is hence guaranteed because a system that was selected to have the discriminatory capacity of detecting a certain external condition may fail to do that. That is, sometimes it fails to fulfil its function of tokening the representational state

whenever the external condition obtains. There are two situations in which there is malfunctioning. First, the producer tokens the representation, but the external condition does not obtain – the representation is false and hence there is misrepresentation. The second malfunctioning case is when the producer does not token the representation even though the external condition obtains – this is just ignorance (or false negative), the external condition is the case but the representation is absent. So, the conclusion is that producer-based teleosemantics does not rule out misrepresentation.

But how exactly can the producer system fail to detect the external condition that it was selected to discriminate and so end up malfunctioning? For instance, how can the frog's visual system fail to fulfil its function of detecting SDMTs? In order to answer this question, two things are required. First, we need to keep in mind the distinction between misrepresentation and ignorance cases since they are not always caused by the same factors. Second, we need to make a distinction between two factors responsible for malfunction – *internal factors* (i.e., something that occurred inside the organism) and *external factors* (i.e., something that occurred outside the organism). After that, I will finish this section by developing an argument to show the plausibility of malfunctioning assignments delivered by producer-based teleosemantics in contrast with the ones delivered by consumer-based teleosemantics.

The standard objection to producer-based teleosemantics is that it fails to leave room for the possibility of misrepresentation cases, not ignorance ones. But in order to have a general picture of malfunctioning cases, it is important to explain how producer-based teleosemantics keeps the room open not only to misrepresentation cases, but also to ignorance cases. Let us start with ignorance cases. So, in virtue of what may the frog's visual system fail to token the representational state despite the presence of the SDMT? Here there are internal and external factors. The most obvious cases of ignorance malfunction due to internal factors are *internal defects*. For instance, in virtue of some pathology the visual system fails to token the

representation despite the presence of the SDMT. The most obvious cases of ignorance malfunction due to external factors are *inappropriate environmental light*. For instance, the system fails to token the representation despite the presence of the SDMT simply because it is embedded in a dark environment. As a result, there is no shadow on the frog's retina or the shadow fails to satisfy the parameters of SDMTs shadows – in both cases, there is no tokening of the representation. This is an external factor because what originates the ignorance – inappropriate light – is something originated outside the frog's organism.[115]

Let us turn to error malfunctions – where the visual system tokens the representation, but the external condition does not obtain. Once again, there are internal and external factors. An internal defect like a neurological damage can lead the visual system to misrepresent the presence of the SDMT. What about the external factors? There are various cases. I will consider two. Consider a machine placed in front of the frog that emits a light ray which goes directly to its retina. It creates a shadow which satisfies the parameters of the SDMT shadow and so it triggers the tokening of the representation. Since the state is representing a SDMT when there is no SDMT around, this is a misrepresentation case. This is an external factor for misrepresentation because its origin is something that occurs outside the organism – the light ray coming from the machine. The second case is a surgical intervention. Suppose that the scientist opens the frog's brain and places an electrode in the right place, resulting in the tokening of the representational state even though there is no SDMT (NEANDER, 2013). This misrepresentation case is caused by an external factor precisely because it is the scientist's surgical intervention that triggers the tokening of the representation despite the absence of any SDMT. Generalizing, anything outside the organism that causes the tokening of the representation despite the absence of the SDMT constitutes an external factor for

[115] Here I am presupposing the biological need of catching SDMTs in the relevant ignorance situations. Sleeping, mating and other situations in which it is conspicuous that there is no such biological need are hence promptly ruled out.

misrepresentation.

So far, so good. There are *internal and external factors* for both misrepresentation and ignorance cases. Producer-based teleosemantics hence does not entail that there are only internal factors for malfunctioning cases. There are external as well as internal factors for the producer system to not fulfil its function of detecting the external condition which it was selected to discriminate. The conclusion is that it is a condition for proper functioning that the producer system is embedded in the right environment.

But here there is a fundamental divergence between proponents of producer-based and consumer-based teleosemantics. The latter claims that the absence of the external correlation between what the producer is capable of discriminating and what is evolutionarily beneficial for the organism is a genuine external factor for the producer system's malfunction. By contrast, proponents of producer-based teleosemantics deny that such correlation is an external factor for malfunctioning. So, consumer-based teleosemantics claims that the holding of the external correlation between SDMTs and nutritive flies is a required external condition for the visual system's proper functioning while producer-based teleosemantics denies that. Some people might worry that the producer-based approach does not allow external factors to lead to malfunction but as we have seen, this is a mistake. Actually, both approaches agree that the malfunctioning of the producer system may be in virtue of internal as well as external factors. The great debate between them is whether it is a genuine external factor for malfunction the *absence of the relevant external correlation* in the environment (e.g., SDMTs and nutritive flies). There is a fundamental clash of intuitions here. How should we assess this debate?

Here it could be argued that the system cannot malfunction in virtue of the absence of the relevant correlation because this is something fully independent of whatever the system does. That is, it has nothing to do with how the system works. However, this argument is flawed from the very start since both teleosemantic approaches agree that a system may malfunction

in virtue of external factors which are also fully independent from the system (e.g., inappropriate environmental light). In what follows, I will develop an argument for producer-based teleosemantics based on the plausibility of malfunctioning assignments to detection systems. This issue plays a crucial role in the assessment of this debate.

It is plausible to claim that the malfunctioning status of the system depends on the surrounding environment, that external conditions are required for its proper functioning. Such plausibility becomes even stronger when this thesis is contrasted with the opposite thesis that proper functioning is just internal proper functioning – which entails that external features play no role in the proper functioning of the system. However, producer-based teleosemantics does not imply this latter thesis. As previously showed, according to it there are external and internal factors in virtue of which a given system may malfunction. But it becomes very implausible to claim that even the producer system with perfect inner proper working under perfect environmental conditions will be malfunctioning if the relevant external correlation simply ceases to hold.

What leads some philosophers to maintain that the malfunctioning status of the system depends on external features is the idea that some external factors are required for the proper functioning of the system. I completely agree with this line of reasoning, but what I am arguing here is that the holding of the relevant external correlation is not among these external factors. The correlation partially explains why a given trait was selected by evolution. The system was selected to detect certain external condition because it is correlated with some other external condition that is beneficial for the system, but this fact doesn't imply that the system was selected to detect this latter condition. That is, the frog's visual system was selected to detect SDMTs because they are correlated with nutritive flies, but this fact doesn't imply that the visual system was selected to detect nutritive flies. The function of the system has to do with how the system internally works and with certain internal and external conditions that are

causally related to its inner working. It has nothing to do with external features to which the system's inner working has no causal relation like external correlations. That is, the proper function of the visual system depends on how it internally works and with internal conditions (e.g., no neurological damage) and external conditions (e.g., appropriate environmental light) that are causally related to its inner working. The proper function of the visual system is not dependent on the correlation between SDMTs and nutritive flies to which the system's inner working has no causal relation. Notice that environmental light is causally related to how the system internally works. Inappropriate light may not trigger the representational state even though the represented external condition obtains (ignorance cases) or it may trigger the representational state despite the absence of the represented external condition (misrepresentation cases).

When one realises that the external correlation has no causal relation with how the producer system internally works and that several external conditions are required for the system's internal proper operation, it becomes very implausible to claim that the absence of the correlation is an external factor for the system's malfunctioning. The realisation of these two things leads to the implausibility of the thesis that the absence of the correlation is an external factor for malfunction. On one hand, there is no causal relation between the external correlation and the system's internal operation – it makes no difference for the tokening of the representation. On the other hand, several external conditions are required for the proper internal operation of the system. The force of the intuition that the absence of the external correlation is responsible for the malfunctioning of the system lies on the plausibility of the idea that the system may malfunction in virtue of external factors. But this intuition is undermined once we recognise that several external factors can be responsible for such malfunctioning. It is implausible to claim that such a weak relation as the external correlation between certain properties in the environment (e.g., SDMTs and nutritive flies) lies among the

external factors responsible for the malfunctioning of the system. By contrast, the plausibility of the claim that some external features are responsible for such a malfunction arises once one realises that these external features are causally related to how the system internally works. The relevant external correlation is a very weak relation to be solely responsible for the malfunctioning of the system in contrast with causal relations.

In sum, the realisation that there are various external factors which lead to malfunction that are causally related to how the system inner works *supresses* the plausibility of claiming that the absence of the correlation is among the genuine external factors for malfunction. The producer system cannot malfunction *simply* in virtue of the absence of the external correlation in the surrounding environment. That said, let's move on and assess the distality problem.

5.4 Functional indeterminacy (II): the distality problem

The core thesis of producer-based teleosemantics is that the content of the representational state is the external condition for which its producer system was selected to discriminate. The producer detects a given distal feature through a causal chain constituted of proximal features – the distal feature causes the proximal ones which ultimately causes the tokening of the representational states. That is, there is a causal chain that that starts with the distal feature, has the proximal features as intermediate causal links and ends with the tokening of the representational state. The question that arises for producer-based teleosemantics is the following. In virtue of what was the producer system selected to discriminate the distal feature but not the proximal one? Since the producer system discriminates the distal feature via the discrimination of the proximal features, why is its function to discriminate the distal feature but not the proximal ones? The threat is that it is indeterminate whether the function of the producer system is to detect proximal or external features and hence that the content delivered by producer-based teleosemantics is indeterminate. This is the distality problem (also called

"the problem of distal content") that is the subject matter of this section.[116]

The distality problem asks why representation R represents the distal feature C but not the proximal feature Q (or vice versa) when Q is a more proximal feature in the causal chain that starts with the presence of C and triggers the tokening of R. That is, which item in the causal chain of proximal and distal features that triggers the tokening of R is the one that determines its content? This problem arises for every producer system in which there is such causal chain. Let's illustrate it with the case of the frog. Let's assume, in light of the debates of the previous sections, that producer-based teleosemantics is the appropriate teleosemantic approach and so that the function of the frog's visual system is to detect SDMTs. There is the following causal chain responsible for the triggering of the frog's representational state R:

SDMT → pattern of light → shadow on the frog's retina → tokening of R

That is, the presence of the SDMT causes a certain pattern of light that causes the shadow on the retina that causes the tokening of the representation. For instance, is the function of the visual system to discriminate SDMTs or appropriate shadows on the retina? It is through the discrimination of the appropriate shadows on the retina that the frog's visual system discriminates SDMTs, so why not say that its function is to discriminate such shadows? It is also indeterminate whether the representation has the proximal content *shadow on the retina* or the distal content *SDMT*. In the latter case the function of the visual system is to detect a certain object in the environment, while in the first case its function is to detect the appropriate shadow on the retina. The truth-conditions of the distal content representation is that a certain external condition obtains in the environment – the presence of the SDMT – while the truth-conditions of the proximal content representation is the presence of an appropriate shadow on the frog's retina.

[116] Fred Dretske was the first to formulate the distality problem for teleosemantics (DRETSKE 1986). Based on it, Ruth Millikan objected that producer-based teleosemantics entails that "all representations must only be of proximal stimuli" (MILLIKAN, 2001, p. 118).

The distality problem is a functional indeterminacy problem that threatens producer-based teleosemantics because it seems that there are equally strong reasons to describe in conflicting ways the function of the frog's visual system. Is its function to detect the proximal stimulus (the shadow on the retina) or the distal stimulus (the presence of the SDMT)? Given that the visual system discriminates the distal feature by discriminating the proximal ones, why give priority to one feature over the others? Finally, its proper functioning status is indeterminate. Consider again the machine that emits the light ray which causes the shadow on the frog's retina and triggers the tokening of the representation. In this case, the visual system is proper functioning provided that its function is to detect shadows on the retina, but it is malfunctioning provided that its function is to detect SDMTs.

The relation between the presence of the SDMT and the shadow on the retina is a causal one. The first causes the later, it is not a mere correlation like the one that holds between SDMTs and nutritive flies. The distality problem is a "horizonal problem" since it asks for the specification of which item in the causal chain that triggers the representation is the one that determines content: SDMT, pattern of light or shadow on the retina? By contrast, "vertical problems" ask for the specification of which description determines content among the competing descriptions of the same item in the causal chain. Which description of the item in the causal chain constituted by the fly determines content – *SDMT* or *nutritive fly*? Note that "shadow on the retina" and "SDMT" concern different items in the causal chain, while "SDMT" and "nutritive fly" are competing descriptions of the same item, namely, the one constituted by the fly. The distality problem asks to locate content among the horizontal axis, not in the vertical axis.[117] That is the reason that the distality problem is a distinctive functional indeterminacy problem. Its solution requires a principled criterion to establish that the

[117] Here I am following Kim Sterelny and Peter Godfrey-Smith in characterizing the distality problem as a "horizontal problem" in contrast with "vertical problems". Cf. STERELNY, 1990; GODFREY-SMITH, 1989.

representational content is constituted by the distal feature but not the proximal one or vice versa. That is, the task is to develop a principled criterion that excludes the proximal features but includes the distal feature in the delivered representational content.

In this section, I will defend that the function of the visual system is to detect SDMTs, not appropriate shadows in the retina or patterns of light, and so that the content of the frog's representation is *SDMT.* In what follows, I will develop a principled criterion which entails that the function of the producer is to discriminate only the distal feature.[118]

Let's start this investigation by excluding two functional assignments located in the extreme links of the relevant causal chain that are easily ruled out by producer-based teleosemantics. The first one lies in its extreme start and gives rise to an overly distal content. Why isn't the function of the producer system to discriminate the light rays before they are reflected on the SDMT? After all, these light rays are the ones that start the causal chain that ultimately creates the shadow on the retina.[119] This functional assignment gives rise to this overly distal content: the representational state represents the light arrays before they reach the SDMT. This functional assignment is easily ruled out simply because the visual system cannot discriminate such light rays. One easily verifies it by just considering the scenario in which such light rays are present but there is no SDMT around – the representational state would not be tokened and so there would be no discrimination (NEANDER, 2017, p. 223). So, it cannot be the function of the visual system to detect such overly distant light rays.

The second functional assignment lies in the opposite side of the causal chain – the extreme end – and gives rise to an overly proximal content. The shadow on the frog's retina triggers a range of retinal firings and neural signs that ultimately token the representational

[118] I have presented the distality problem as a problem for producer-based teleosemantics since this is the theory that I am developing here and because it is "particularly problematic" for producer-based teleosemantics (RYDER, 2009, p. 263). But this is a general problem for teleosemantics and also to causal theories of content that try to determine content in terms of the causal relations of the representational state.

[119] This objection was originally developed by Carolyn Price to Neander's producer-based teleosemantic theory, cf. NEANDER, 2017, p. 281.

system. So why isn't the visual system's function to detect such retinal firings or neural signals? The result would be the assignment of this overly proximal content: the representational state represents such retinal firings and neural signs. How to rule out such functional assignment?

Neander has defended the following criterion to rule out neural signals and retinal firings as candidates for detection function to the visual system. She claims that "these are already eliminated, because they are part of the response [i.e. detection] involved in the relevant response function [i.e., detection function]" (NEANDER, 2017, p. 222). That is, the neural signals and retinal firings constitute the very detection process that is the function of the producer system to fulfil – to token the representational state in response to the presence of the relevant external condition. Since the representational state cannot represent elements in the detection process responsible for its very tokening, it follows that it cannot be the function of the producer system to detect retinal firing or neural signs and so that they are not constitutive of the representational content.

However, there is a fundamental problem with this criterion that threats its success in ruling out that the function of the visual system is to detect retinal firings or neural signs. In the centre of this criterion lies the assumption that the detection process is constituted by retinal firings, neural signals, etc. But this is not mandatory. The room is still open for the assumption that the detection process encompasses neural signs and other elements but excludes retinal firings. It seems that it is arbitrary to assume that both retinal firings and neural signals are constitutive of the detection process. After all, maybe it is constituted by neural signals and other internal elements because its function is precisely to detect retinal firings. Why not? Here there is the risk of presupposing that retinal firings are constitutive of the detection process just to rule out the assignment that the function of the visual system is to detect retinal firings. Unless some principled justification is available to include both retinal firings and neural signals among the items that constitute the detection process, it is arbitrary to simply assume

that this is the case. The conclusion is that in the absence of such justification, this criterion fails to rule out retinal firings as a candidate for what is the function of the system to detect. Rather, it keeps the door open for it.

But I think that there is a way of saving this criterion by providing the following justification for the claim that both retinal firings and neural signs are constitutive of the detection process, and so for ruling them out of what is the function of the visual system to detect. This is a consequence of the lower-level thesis and the functional analysis previously proposed to specify the producer system's function. Let's show it in detail. The lower-level thesis claims that the function of the producer system is its direct effect and that the functional analysis is required precisely to specify which one is the system's direct effect. Accordingly, the direct effect of the producer system is the effect that appears in the lowest level of the functional analysis of the organism in which the producer system appears *as an unanalysed component*. However, what happens is that such neural signals and retinal firings are themselves components of the relevant producer system in the case of the frog – the visual system. They appear in the functional analysis only at the level in which the visual system is itself analysed. Retinal firings and neural signals are unequivocally components of the visual system. At the level in which the visual system is an unanalysed component, they do not appear at all. Hence, it cannot be its direct effect to detect neural signals or retinal firings and so they cannot be what is the function of the visual system to detect.

The conclusion is that it is not the function of the producer system to detect either overly distal features like the light rays before they are reflected from the SDM, or overly proximal features like retinal firings or neural signals. Its function is to detect something between these extremes – the SDMT, the light pattern before it reaches the retina or the shadow on the retina. But how to develop a principled criterion to settle this issue?

The distality principle

What is required is a distality principle that shows that the function of the producer system is to discriminate the distal feature, not the proximal ones. While proper functioning, the producer system discriminates distal feature C via the discrimination of the proximal feature Q. Why does the producer system have the function of discriminating distal feature C rather than proximal feature Q? It is perspicuous that the producer system was designed to detect both C and Q, after all the detection of C is achieved via the detection of Q. So, the question is not whether when proper functioning the system detects C or Q, after all it detects both features. Rather, the question is whether the system was *selected* for the detection of C or Q. That is, is it the function of the producer system to detect the distal or proximal feature?

The principle that I will defend here was previously mentioned in the assessment of the constancy mechanism proposal for the limits of intentionality in the third chapter. The function of the producer system is to discriminate the distal but not the proximal feature because there is a variety of proximal stimuli coming from the same distal feature that triggers the tokening of the representational state. That is, the producer system was selected to have the capacity of bridging across different proximal stimuli coming from the same distal feature in order to detect the distal feature. So, the producer system discriminates the distal feature via different stimuli routes – not only via the same proximal stimulus.

> DISTALITY PRINCIPLE: the producer system has the function of detecting distal feature C but not proximal stimulus Q in the causal chain that triggers the tokening of representational state R because there is a multiplicity of intermediate proximal stimuli that triggers the tokening of R in response to the presence of C.[120]

[120] Fred Dretske was the first to formulate this principle, cf. DRETSKE, 1981, p. 163; 1986. Kim Sterelny, Peter Schulte and Justin Garson have also defended varieties of this principle, STERELNY, 1990; SCHULTE, 2018; GARSON, 2018.

Such multiplicity of stimuli routes responsible for the triggering of the representational state rules out that the function of the producer system is to detect the intermediate features between the sensory apparatus and the distal feature. The reason is that there is no single intermediate feature that is present in all causal chains. Sometimes the token of the representational state is triggered by proximal stimulus Q_1, while other times it is triggered by Q_2, or Q_3 ... or Q_n. The only external feature that is present in all these causal chains is precisely the external feature C and hence the producer system may detect the presence of C without always detecting the presence of any specific proximal stimulus Q_n. Evidently, such detection will always be via the detection of some proximal stimulus, but which one is the present proximal stimuli varies from situation to situation. The presence of C is the only normal cause that triggers the tokening of the representational state understood as the causal link that is always present in the causal chain in biological normal situations (DRETSKE, 1986, p. 168-9). So, the conclusion that the function of the visual system is to detect C.

It is arbitrary to assume that a given organism which is capable of detecting the distal feature only via a single proximal stimulus has the function of detecting the distal feature but not this proximal stimulus. However, things change when the organism is capable of detecting the distal feature via more than one stimulus route. In this case, there is a principled reason to claim that the organism was designed to detect the distal feature, not any individual proximal stimulus – the distal feature is the only item present in all distinct causal chains that trigger the tokening of the representational state. Suppose that R is triggered by two distinct stimuli routes coming from distal feature C – via Q_1 or via Q_2. Q_1 is sufficient to trigger R and so is Q_2. Thus, it is arbitrary to claim that the function of the producer system is to detect Q_1 because it would be equally plausible that its function is to detect Q_2 – why privilege one stimuli route rather than the other? The only non-arbitrary functional assignment is that the function of the producer

system is to detect C because it is the only external feature that is present in both stimuli routes. Hence, the conclusion is that R represents C, not Q_1 or Q_2.

Let's illustrate this distality principle. The frog's visual system has the function of detecting the SDMT by tokening the representational state but not the shadow on the retina or the light pattern. That is the case because it is capable of detecting SDMTs via a variety of light patterns and shadows on the retina that that triggers the tokening of the representational state. So, the content of the representation is *SDMT*. The tokening of the representation is triggered by a multiplicity of retinal stimulation patterns and light patterns in different circumstances.[121]

Here an elucidation is necessary. Does this distality principle amount to the requirement that producer systems should employ constancy mechanisms in the production of representational states in order for the representations to have distal contents? No. What happens is that the employment of the constancy mechanism is just a special case of the satisfaction of the above distality principle – it is not the only way in which the producer system may satisfy the distality principle. That is, it is just one among other ways of implementing the many-one mapping function from the multiplicity of proximal stimuli to the tokening of the representational state. Let me explain why. As I have been using "constancy mechanism", the employment of the constancy mechanism occurs only in the production of representational states within a single-sense modality. That is, constancy mechanisms are employed only in the production of single sensory representational states – visual representations, auditory representations, etc. Notice that the previous examples of the employment of constancy mechanism resulted in the production of single sensory visual representations – the honeybee's, vervet monkey's and the frog's visual representational states. However, the implementation of the many-one mapping function required for the satisfaction of the distality principle is also

[121] It is interesting to note that the visual system of the toad, an organism that has a very similar visual system with that of the frog, exhibit size constancy. It can detect the same distal object despite the size distinction of retinal images that are produced on the retina. The retinal image size varies with the distance that the object is from the retina. The toad detects the real size of objects, not the size of retinal images. Cf. EWERT, 1980, p. 74.

achievable by producer systems equipped with two or more sense modalities that produce multimodal representational states (DRETSKE, 1986, p. 131). For instance, you may represent the presence of a dog either via the visual sense when you see the dog or via the auditory sense when you hear it barking. So, in order to highlight this difference, I will use "constancy mechanism" just to refer to the specific implementation of this general model by a single-sense representational state.

It is not surprising why I have appealed only to constancy mechanisms in the debate on the minimal conditions for intentionality. After all, multimodal representations have a far more complex intentionality than single-modal representational states and hence are not within the scope of the debate on the limits of intentionality – it is very implausible to cast doubt on the intentional status of systems that produce multimodal sensory states. The producer system should be sufficiently complex to produce representations via two or more sense modalities.

I think that this shows that the distality principle is not only fully compatible with the theory of mental representation that has been developed in this thesis but is also motivated by it. This theory defends the dual proposal for the limits of intentionality according to which the employment of the constancy mechanism in the production of the primitive states is a minimal condition for intentionality. It follows that the primitive sensory states considered by this theory as genuinely intentional satisfy the distality principle. This constitutes a reason for the principled character of the distality principle in light of the theory of mental representation developed in this thesis. The distality principle is not an ad hoc or arbitrary criterion proposed only to give a response to the distality problem.

The disjunction objection

So far, so good. But the distality principle is chased by the following threat (DRETSKE, 1986). It is not true that the distal feature is the only external feature that is present throughout

all stimuli routes that trigger the tokening of the representation. The disjunction objection consists in claiming that there is a proximal item that is also always present and thus constitutes a normal cause for the triggering of the representation. Namely, the disjunction of all proximal stimuli that triggers the representational state, $\{Q_1 \text{ v } Q_2 \text{ v } \ldots \text{ v } Q_n\}$. Such a disjunctive proximal feature is always present in the causal chain just like distal feature C. They are both equally normal causes for the tokening of the state. Hence, it is arbitrary to claim that it is the function of the producer system to detect C but not the disjunctive proximal feature $\{Q_1 \text{ v } Q_2 \text{ v } \ldots \text{ v } Q_n\}$ and so that that the representational state represents only C. No matter how many stimuli routes the producer system may have access in order to respond to the distal feature, there will still be the possibility of describing its function as the detection of the disjunction of the proximal stimuli that trigger the representation.

This is a very serious objection to the viability of the distality principle solution to the distality problem. How to assess it? What is the reason that the function of the producer system is to detect C but not the disjunction of proximal stimuli $\{Q_1 \text{ v } Q_2 \text{ v } \ldots \text{ v } Q_n\}$? It is of no help to claim that its function is to detect C because $\{Q_1 \text{ v } Q_2 \text{ v } \ldots \text{ v } Q_n\}$ is a disjunctive feature, arguing that producer systems could not have been selected to detect disjunctive features. It is ad hoc to just stipulate that.[122] In what follows, I will develop a response to the disjunction objection that takes a different route.

The producer system has the function of detecting the distal feature but not the proximal disjunctive feature because the detection of the latter is just a means to achieve the detection of the distal feature. That is, the adaptive effect for which the producer system was selected is the detection of the distal feature. The detection of the proximal disjunctive feature is just a means

[122] Dretske defended that the only way to guarantee that the state represents the distal but not the disjunctive proximal feature is to require from the organism the capacity of some form of associative learning (DRETSKE, 1986, pp. 170-1). But I will not assess Dretske's response to the disjunction objection here – it strikes me as unmotivated since it is restricted to representations of organisms capable of associative learning. Here I will develop a solution to this problem that is not restricted to organisms with such associative learning capacity.

to achieve this end. It strikes me that this is the right intuition that should give rise to the response to the disjunction objection, but how to justify it? If there is no available justification, this response is arbitrary.

That is the case in virtue of the disjunctive character of this proximal feature. The only proximal feature present in all stimuli chains responsible for the tokening of the state is disjunctive precisely because there are different means to detect the distal feature for which the system was selected to detect. Sometimes the detection of C is via Q_1, other times via Q_2, etc. But why should one accept such a conclusion? The argument for it arises from the assessment of *alternative evolutionary environments* for the organism. The organism was designed in the original environment to respond to C via the detection of disjunctive proximal feature $\{Q_1 \text{ v } Q_2 \text{ v } Q_3\}$. Now suppose that the organism has evolved in an alternative environment that is very similar to the original one. The only difference is that Q_3 is absent and that another proximal stimulus Q_4 is present, such that now the detection of Q_4 is also a means for the detection of C. But the detection of Q_3 is no longer a means to detect C, in this alternative environment it is the detection of the disjunctive proximal environment $\{Q_1 \text{ v } Q_2 \text{ v } Q_4\}$ that leads to the detection of C, not $\{Q_1 \text{ v } Q_2 \text{ v } Q_3\}$. Finally, suppose that Q_3 and Q_4 are very similar proximal stimuli (e.g., two retina images with different but very similar sizes). What would have happened in this alternative environment? How would the organism have evolved?

It is perspicuous that evolution would have selected the producer system to be causally sensitive not to $\{Q_1 \text{ v } Q_2 \text{ v } Q_3\}$ as it happened in the original scenario, but to $\{Q_1 \text{ v } Q_2 \text{ v } Q_4\}$. Why is that the case? Just like in the original scenario $\{Q_1 \text{ v } Q_2 \text{ v } Q_3\}$ is the disjunctive proximal feature that is a means to achieve the adaptive effect of detecting C, in the alternative scenario it is the detection of $\{Q_1 \text{ v } Q_2 \text{ v } Q_4\}$ that leads to the adaptive effect of detecting C. Notice that it could not be the function of the producer system to detect the disjunctive proximal feature $\{Q_1 \text{ v } Q_2 \text{ v } Q_3 \text{ v } Q_4\}$. The reason is that there are other alternative scenarios in which both Q_3

and Q_4 are absent and instead another proximal stimulus Q_n is a means for the detection of C, such that in these other alternative scenarios it is the detection of $\{Q_1 \text{ v } Q_2 \text{ v } Q_n\}$ that leads to the detection of C. The conclusion is that there is no disjunctive proximal feature $\{Q_1 \text{ v } Q_2 \text{ v } ... \text{ v } Q_n\}$ that is always present, throughout all actual and alternative evolutionary environments, in the stimuli chain responsible for the tokening of the representation. C is the only feature that is always present. So, there is no disjunctive proximal feature that constitutes a normal cause for the tokening of the representational state throughout all actual and alternative evolutionary environments. This fact shows that the disjunctive character of the disjunctive proximal feature is the reason that the detection of this feature is just a means for the detection of the adaptive distal feature C.

In sum, the assessment of alternative scenarios show that evolution would have selected the organism's producer system in order to discriminate distinct disjunctive proximal features. The reason is that it changes from scenario to scenario which one is the relevant disjunctive proximal feature whose detection leads to the detection of the adaptive distal feature. Hence, it cannot be the function of the producer system to detect the disjunctive proximal feature in the original evolutionary environment. The distal feature is the only feature that is present in all these scenarios and so the conclusion is that it is the function of the producer system to detect the distal feature, not any disjunctive proximal feature. The representational state has a distal content, not a proximal one.

Let's contrast my response to the disjunction objection and Neander's solution for the distality problem. Neander rejects the above distality principle according to which the producer system has the function of detecting the distal feature, not the proximal ones, because there is a multiplicity of intermediate proximal stimuli that triggers the tokening of the representation. Instead, she proposes the following principle: "a sensory system is only adapted to respond to the more proximal items because doing so is the means by which it responds to the more distal

ones, and not vice versa" (NEANDER, 2013, p. 34). This principle is very different from the distality principle. However, it has a certain similarity with my response to the disjunction objection.

Neander's principle strikes me as ad hoc because it is hard to conceive a principled argument based on the producer-based teleosemantics to justify this principle as a solution for the distality problem. However, things change when one's preferred solution for the distality problem is the previously defended distality principle. As I have tried to show, the distality principle is not ad hoc or arbitrary. There is a fundamental argument for it which is based on the multiplicity of intermediate stimuli routes that trigger the tokening of the representation. Furthermore, there is a strong motivation for the distality principle based on the dual proposal for the limits of intentionality that I have defended in the fourth chapter. However, the distality principle is threaten by the disjunction objection – why is it the function of the producer system to detect the distal feature rather than the disjunctive proximal feature?

My strategy to give a response to this objection consists by arguing that the function of the producer system is to detect the distal feature because the detection of the disjunctive proximal feature is just a means for the detection of the adaptive distal feature. But how to justify it? This move is plainly justified as a response for the disjunction objection because it is in virtue of *the disjunctive character* of the relevant disjunctive proximal feature that the detection of this feature is just a means for the detection of the distal feature. The argument for it arises in *the assessment of alternative evolutionary* environments for the organism. Evolution would have selected the producer system to discriminate distinct disjunctive proximal features in alternative evolutionary scenarios. The reason is that *it changes from scenario to scenario* the relevant disjunctive proximal feature whose detection leads to the detection of the adaptive distal feature.

Finally, it is important to notice the contrast of plausibility between two claims. It is

more plausible to claim that the detection of the disjunctive proximal feature is a means to the detection of the distal feature (my response to the disjunction objection) than to claim that the detection of a non-disjunctive proximal feature in a specific stimuli chain is just a means to the detection of the distal feature (Neander's claim). The reason for such contrast is *the disjunctive character* of the former proximal feature. Such disjunctiveness is what grounds the claim that the disjunctive proximal feature is just a means to the distal feature. That is the case because different stimuli routes lead to the same distal feature *throughout alternative evolutionary scenarios*. This is the fundamental difference. But since Neander, in her solution for the distality problem, does not appeal to the multiplicity of stimuli routes that give rise to the relevant disjunctive proximal feature, her principle lacks the required justification. As a result, it looks ad hoc.

Let's finish this section by assessing an objection that Carolyn Price has developed to Neander's solution for the distality problem that also applies to my response for the disjunction objection (PRICE, 2014, p. 590). My response to the disjunction objection implies that the function of the frog's visual system is to detect SDMTs, not the disjunctive proximal feature constituted by the shadows on the retina that triggers the tokening of the representation. After all, the detection of shadows on the retina is just a means to achieve the adaptive detection of SDMTs. But here it could be objected that in the end such line of reasoning will lead to the claim that the frog's visual system represents nutritive flies, not SDMTs. The reason is that the assessment of alternative evolutionary environments shows that evolutionary selection would have select the visual system that produces representations that track nutritive flies, not SDMTs. So, why couldn't my own preferred producer-based function assignment be ruled out for the same reason that I have ruled out the assignment that the function of the producer system is to detect the disjunctive proximal feature? After all, the visual system was selected to discriminate SDMTs only because it was by this means that it was capable of discriminating

nutritive flies. Hence, the function of the visual system is to detect nutritive flies, not SDMTs.[123]

The problem with this objection is that it misses what is the issue behind the distality problem. As previously noted, the distality problem is a *horizonal problem*. The issue of the distality problem (and the disjunction objection) is not how to properly describe the distal feature that the visual system is supposed to detect – as a SDMT or as a nutritive fly. This is the issue of vertical problems. Rather, the issue of the distality problem is the specification of the item in the causal chain that the producer system has the function to detect. Is it the distal feature or some proximal feature? How to properly describe the distal feature is a further question. The distality problem does not ask what the appropriate description of the distal feature among the competing descriptions is – "SDMT", "nutritive fly", etc. Remember that in the beginning of the debate on the distality problem, I have assumed that the proper description of the distal feature is SDMT (not nutritive flies) in light of the arguments that I have developed for producer-based teleosemantics in the former sections of this chapter. But what is at issue is not this or that description of the distal feature. Rather, the issue is whether the function of the visual system is to detect the distal feature or some proximal feature in the first place.

However, now it could be replied that nothing prevents the opponent of producer-based teleosemantics to argue that Price's objection should be withdrawn from the context of the debate on the distality problem in order to turn it into a general objection to producer-based teleosemantics. That is, the general objection that the producer system was selected for its capacity to discriminate the sensible properties of the relevant distal feature (e.g., SDMT) only because it was by that means that it was able to discriminate the evolutionary beneficial properties of the distal feature (e.g., nutrition).

It strikes me that this reason should *only be applied to the vertical problem*, not to the

[123] Neander assesses Price's objection in her last published book, but it is not clear at all that her response rebuts this objection, cf. NEANDER, 2017, p. 223. In what follows, I will take a different route.

horizontal ones. That is, it should only be applied for the specification of the item in the causal chain that is the function of the producer system to discriminate. But if one rejects this claim and really wants to turn Price's objection into a general objection to producer-based teleosemantics, one should then consider the general picture of the debate between producer-based and consumer-based teleosemantic approaches. It is not enough now to consider only the debate on the distality problem. As I have showed in the previous sections, there are strong arguments in favour of producer-based teleosemantics that should be considered when this global debate between producer-based and consumer-based teleosemantics arises. First, I have argued in the second section that producer-based teleosemantics render content determinate in light of the concertina and relativity problems, while there is no prospect of consumer-based teleosemantics rendering content determinate relatively to the relativity problem. Second, I have argued in the third section that producer-based teleosemantics delivers plausible malfunctioning assignments to detection systems, by contrast with consumer-based teleosemantics. In light of these strong reasons for producer-based teleosemantics, it is hard to argue that this global version of Price's objection will make the balance leans towards the rejection of producer-based teleosemantics. Notice that the debate on the disjunction objection is a local issue that cannot *by itself* decide the general debate between producer and consumer-based teleosemantics. In the big picture, there are strong reasons in favour of the former. So, my conclusion is that the balance leans in favour of producer-based teleosemantics, not against it.

5.5 The source of error objection

In this final section, I will assess one last objection to producer-based teleosemantics. Consider a situation in which the producer, consumer and all other frog's systems are properly functioning and in which the visual system detects a non-nutritive SDMT. As a result, there is

no increase of fitness. But how is it possible that there is no increase of fitness even though all systems of the organism are proper functioning? This is the source of error objection. Its assessment comes with an important lesson for teleosemantics.

This situation is perspicuously not problematic for consumer-based teleosemantics because it claims that the function of the representational state is to detect nutritive flies and that the function of the consumer systems is to catch and digest nutritive flies. So, it follows that in the above situation the consumer systems are malfunctioning since they failed to catch and digest nutritive flies and hence there is no incompatibility between the non-increasing of fitness in this situation and consumer-based teleosemantics. By contrast, this situation is problematic for producer-based teleosemantics. As previously argued, this approach claims that the function of the producer system is to detect SDMTs and that the function of the consumer systems is to catch and digest SDMTs. So, in the above situation even though the producer and consumer systems perform their functions of respectively detecting, catching and digesting SDMTs, there will be no increase of fitness because the relevant SDMT is not nutritive. The fact that, according to producer-based teleosemantics, there are cases in which all systems are proper functioning but still there is no increase of fitness seems to be incompatible with the aetiological conception of biological function and with the very spirit of teleosemantics. After all, if all systems are proper functioning but there is no increase of fitness, where is the source of error responsible for the nonincreasing of fitness? The organism is composed of a variety of systems, it is plausible to assume that there is increase of fitness provided that all systems are properly functioning. However, such assumption is wrong. In this final section, I assess the situation in which, according to producer-based teleosemantics, there is no increase of fitness despite the proper functioning of all systems, and I show why this is not problematic and hence that the source of error objection is flawed.[124]

[124] This objection was proposed by Prof. Matthew Parrott in conversation.

In this situation there is no increase of fitness even though all organism's systems are properly functioning in virtue of some external feature of the surrounding environment. That is, in this situation *the source of error lies in the environment*. To guarantee the increase of fitness, it is not only required that the biological systems of the organism are proper functioning, but also that the organism is embedded in an appropriate environment. When the surrounding environment is inappropriate, there is no guarantee of the increasing of fitness. This is the lesson to be taken from this situation. Let's illustrate it with two variations of this situation.

Suppose that the scientist fools the frog by placing a SDMT in front of it – a pellet. The frog's consumer and producer systems are proper functioning and so the visual system will detect it, the motor system will catch it, the digestive system will digest it, etc. However, in the end there will be no increase of fitness because the pellet is non-nutritious. But if the frog's consumer and producer systems are properly functioning, where is the source of error? It must lie somewhere. In this case, the source of error is the scientist fooling the frog. It is important to notice that evolution always operates in a specific environment, the favoured traits are always selected against a background environment. Since evolution is a selection process which operates throughout the history of the species, the relevant environment is the historical environment of the species. So, the species was not designed by evolution to function in any environment whatsoever, but to function in its historical environment. However, in this case the relevant environment is not the frog's historical environment – the scientist changed it insofar as a pellet was placed in front of the frog. That is the reason that the error lies in the environment. If the background environment were the frog's historical environment, there would be increase of fitness. The conclusion is that the fact that all systems are proper functioning is not a guarantee of increasing of fitness. Rather, it is guaranteed only when (1) the producer and consumer systems are proper functioning and (2) the surrounding

environment is the organism's historical one. If one of these two conditions fail to obtain, there is no guarantee of increase of fitness.[125]

Now let's consider a second variation of the situation in which the biological systems of the organism are properly functioning but there is no increase of fitness. Suppose that the frog's consumer and producer systems are proper functioning but there is no perfect correlation in the frog's historical environment between SDMTs and nutritious flies. That is, there are non-nutritious SDMTs in this environment. Now let's suppose that the following correlation holds: nine out of ten SDMTs are nutritious. Hence, in one out of ten cases in which the frog's visual system detects a small-dark-moving thing, it is not detecting a nutritious fly and so there is no increase of fitness even though all frog's systems are proper functioning. But if the relevant environment is the frog's historical environment and the frog's producer and consumer systems are proper functioning, where is the source of error? The fact is that assuming the producer-based approach, in one out of ten cases there is no increase of fitness even though all systems are proper functioning and the relevant environment is the frog's historical one.

Here the source of error lies again in the environment because this specific environment is inappropriate for the frog. There is no increase of fitness because in that specific environment, SDMTs are not nutritive flies. This second case gives rise to an important lesson for teleosemantics. There are a plenty of cases in which a trait is selected even though the relevant correlation *is not perfect*. What matters for evolutionary selection is that the bonus of having the trait is higher than the onus of having it. The selection of a trait by evolution is always a matter of the balance pending for its favouring, not for its disfavouring. In this second case, evolution selected the visual system because even though it will have the effect of detecting food in only nine out of ten cases, the bonus of having it (i.e., the obtainment of food

[125] While presenting the functional indeterminacy problem, Dretske originally wondered whether the error lies in the external environment when the relevant correlation is broken. Cf. DRETSKE, 1986, pp. 166-8.

in nine out of ten cases) is higher than the onus of having it (i.e., the loss of energy in one of out of ten cases). Therefore, it is true that there is no increase of fitness in the specific occasion in which the visual system detects a non-nutritious thing, but it is also true that this trait is evolutionary beneficial when globally considered. That is, in this occasion the trait is not locally beneficial, but it is globally beneficial because there is a general increase of fitness. The lesson is that the increasing of fitness resulted of the performance of the function of a trait should be always considered globally, not locally.

The difference between these two cases is just that in the first one the relevant external feature is the scientist fooling the frog, while in the second case the relevant external feature is the imperfect correlation between SDMTs and nutritive flies in the frog's historical environment. The fact that the biological systems of the organism are proper functioning and that the surrounding environment is the organism's historic environment guarantees global increased fitness, not a local increase. The local increase of fitness requires a surrounding environment in which SDMTs are nutritive flies. The scientist intervention and the minority cases in which SDMTs are not nutritive flies (in virtue of imperfect correlation) are both cases in which there is no increasing of fitness because these local environments are inappropriate. It is a general feature of evolutionary selection that the designed organisms will probably be in inappropriate environments and as a result there will be no increasing of fitness even though the biological systems of the organisms are functioning properly. The conclusion is that the situation in which the organism's systems are functioning properly but there is no increase of fitness is not problematic for producer-based teleosemantics. Rather, the situation is fully compatible with it. The source of error objection to producer-based teleosemantic is hence flawed.

Conclusion

I finish this chapter with a contrast between the theory of representational content here developed and the dual proposal for the minimal conditions for intentionality developed in the fourth chapter. The dual proposal is partly constituted by the success pattern condition which requires the presence of a success pattern in the behavioural output of the organism. This condition focus on the success conditions of the behavioural output produced by the consumer system to establish a minimal condition for intentionality. This condition is hence output-oriented. Thus, the dual proposal for the minimal conditions for intentionality is (partially) *output-oriented*.[126] By contrast, the producer-based teleosemantic theory developed in this chapter is *input-oriented* since it focuses on the function of the producer system in order to determine the content of the representational state. At this point it could be objected that there is an inconsistence (or incoherence) between my approach to representational content and my approach to the problem of demarcation – the former is *output-oriented* while the latter is *input-oriented*. But should not a theory of mental representation be either input-oriented or output-oriented? Is it possible for a theory to be input-oriented regarding representational content and output-oriented regarding minimal conditions for intentionality?

In fact, there is no inconsistency at all. It is plainly possible for a theory of mental representation to have these input-oriented and output-oriented approaches. The reason is that a theory of content is distinct from a theory of representational status. A theory of content is supposed to give an account of *the content problem* ("in virtue of what is a given state a representational state?"), while a theory of representational status is supposed to give an account of *the representational status problem* ("provided that a given state is representational, in virtue of what does it represents this state but not another state?"). Notice that the problem

[126] The dual proposal is partially output-oriented as well because it is also constituted by the constancy mechanism condition which is input-oriented, as argued in the fourth chapter. However, in what follows I set side this input-oriented aspect of the dual proposal and focus on its output-oriented aspect.

of demarcation is a problem for the teleosemantic solution to the representational status problem, while producer-based teleosemantics is a teleosemantic solution to the content problem. There is no incoherence in assuming an input-oriented approach to the content problem – producer-based teleosemantics – and an output-oriented approach to the representational status problem – the dual proposal.

In this final chapter, I developed and defended the producer-based teleosemantic approach to representational content. The content of the representational state is determined by the function of the producer system, the effect for which the producer system was selected, not by the consumer system's function. The producer system was selected to have the capacity to discriminate a certain external condition and it is this external condition that constitutes representational content. I assessed two functional indeterminacy problems that threaten the viability of producer-based teleosemantics – the concertina and distality problems – and I showed how to develop this approach in order to give an account of these problems and in doing so to determine content. I have also developed a general defence of producer-based teleosemantics based on the plausibility of malfunctioning assignments to detection systems. Finally, in this last section I have showed why the source of error objection does not establish a threat for the viability of this teleosemantic approach. My conclusion is that producer-based teleosemantics is the most plausible account of the content of primitive representational states.

CONCLUSION

In this thesis, I have developed and defended a teleological theory of mental representation. I tried to show that it succeeds in giving an account of primitive representational states, at least considering the previously assessed problems and objections. However, there are several problems and objections to teleosemantics that were not assessed. They are not in the scope of this thesis. For example, I have said nothing about the well-known swampman objection to teleosemantics.

Another issue is that in this thesis I have throughout assumed the aetiological conception of biological function without exploring the virtues of alternatives accounts of biological function. Presumably, it is not possible to give an account of all representational states based solely on the aetiological conception. Maybe it is required to appeal to other conceptions of biological function. That is, there is no unique conception of biological function based on which one can explain the nature of all representational states and so it is necessary to appeal to alternative accounts. It is my hope that the teleological theory defended here is plainly compatible with such a pluralist view in such a way that this theory may be expanded to embrace non-aetiological conceptions of biological functions.

Finally, it is not clear how this teleological theory would best be developed to give an account of more sophisticated representational states like conceptual and personal ones. This is probably the greatest challenge facing not only teleological theories, but naturalist theories of mental representation in general. It strikes me that the problem of how to expand these naturalist theories in order to give an account of more complex representational states should guide the next developments of naturalist theories in the years to come. Let us wait and see.

The teleological theory here developed is a powerful framework that provides a naturalist account of primitive mental representations. Even if it turns out that one cannot

develop teleosemantics to give an account of more sophisticated representational states, its value is still guaranteed in light of its account of primitive representations. This is by itself a great achievement.

REFERENCES

ANTONY, L. & LEVINE, J. 1991. *The nomic and the robust*. In “Meaning in mind – Fodor and his critics” (B. LOEWER & G. REY, ed.). Blackwell. Oxford & Cambridge, MA., pp. 1-16.

AGAR, N. 1993. *What frogs really believe?*. Australasian Journal of Philosophy, 71, 1, pp. 1-12.

ARTIGA, M. 2016. *Liberal representationalism: a deflationist defense*. Dialectica, 70, 3, pp. 407-30.

BATES, E. & L. CAMAIONI & V. VOLTERRA. 1975. *The acquisition of performatives prior to speech.* Merrill-Palmer Quarterly of Behavior and Development, 21, 3, pp. 205-26.

BECKERMANN, A. 1988. *Why tropistic systems are not genuine intentional systems*. Erkenntnis, 29, 1, p. 125-42.

BICKLE, J. 2013. *Multiple realizability*. The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.), URL = < https://plato.stanford.edu/entries/multiple-realizability/>.

BLAKEMORE, R. 1975. *Magnetotactic bacteria*. Science, 190, pp. 377-9.

BLOCK, N. 1986. *Advertisement for a semantics for psychology*. Midwest Studies in Philosophy, 10, 1, pp. 615-78.

—— 1990. *Can the mind change the world?.* In "Meaning and method: essays in honor of Hilary Putnam" (BOOLOS, ed.), Cambridge University Press, Cambridge, pp. 137-70.

BRUSH, S. 1976. *The kind of motion we called heat: a history of the kinetic theory of gases in the 19th century*. *Book 1 – Physics and the atomists*. North-Holland Publishing Company. Amsterdam, New York & Oxford.

BRENTANO, F. 1874 [1995]. *Psychology from an empirical standpoint*. Routledge and Kegan Paul. London.

BURGE, T. 1993. *Mind-body causation and explanatory practice*. In “Mental causation” (HEIL, J. & MELE, A., eds.), Clarendon Press, Oxford.

—— 2010. *Origins of objectivity*. Oxford University Press. Oxford.

CALL, J. & TOMASELLO, M. 2007. *The gestural communication of apes and monkeys.* Lawrence Erlbaum. Mahwah, New Jersey.

CHURCHLAND, P. M. 1981. *Eliminative materialism and the propositional attitudes*. The Journal of Philosophy, 78, 2, pp. 67-90.

COHEN, J. 2015. *Perceptual constancy*. In "The Oxford handbook of philosophy of perception" (M MATTHEN, ed.), Oxford University Press. Oxford, pp. 621-39.

CRANE, T. 2016. *The Mechanical mind*. Routledge. New York.

CUMMINS, R. 1975. *Functional analysis*. Journal of Philosophy, 72, p. 741-765.

DANIELS, N. 2016. *Reflective equilibrium*. In The Stanford Encyclopedia of Philosophy (ZALTA, E., ed.). URL = < https://plato.stanford.edu/entries/reflective-equilibrium/>.

DENNETT, D. 1987. *The intentional stance.* MIT Press. Cambridge, MA.

—— *The myth of original intentionality*. In K.A. Mohyeldin Said, W.H. Newton-Smith, R. Viale, and K.V. Wilkes (eds.), "Modelling the Mind". Oxford: Clarendon Press, pp. 43-62.

—— 1991. *Real patterns*. The Journal of Philosophy, 88, 1, pp. 27-51.

DRETSKE, F. 1981. *Knowledge and the flow of information*. MIT Press. Cambridge, MA.

—— 1986. *Misrepresentation.* In "Mental Representation" (STICH & WARFIELD, eds.), Basil Blackwell, Oxford & Cambridge, MA.

—— 1988. *Explaining behaviour*. MIT Press. Cambridge, MA.

—— 1995. *Naturalizing the mind.* MIT Press. Cambridge, MA.

EWERT, J. P. 1980. *Neuroethology: an introduction to the neurethological fundamentals of behavior*. Springer Verlag. Berlin.

FODOR, J. 1986. *Why paramecia don't have mental representations.* Midwest Studies in Philosophy, 10, pp. 3-24.

—— 1987. *Psychosemantics*. MIT Press. Cambridge, MA.

—— 1990. *A Theory of content and other essays*. MIT Press. Cambridge, MA.

—— 1991. *Replies*. In "Meaning in mind – Fodor and his critics" (B. LOEWER & G. REY, ed.). Blackwell, Oxford, UK & Cambridge, MA, p. 255-319.

—— 1998. *Concepts - Where cognitive science went wrong*. Oxford University Press. Oxford.

GARSON, J. 2018. *Do constancy mechanisms save distal content?* URL = < https://www.justingarson.com/papers/>.

GODFREY-SMITH, 1989. *Misinformation*. Canadian Journal of Philosophy. 19, 4, pp. 533-550.

GOODMAN, N. 1979. *Fact, fiction and forecast*. Harvard University Press. Cambridge, MA and London, UK.

HILBERT, D. R. 2005. *Color constancy and the complexity of color*. Philsoophical Topics, 33, pp.141-58.

JACOB, P. 2007. *What minds can do – intentionality in a non-intentional world.* Cambridge University Press. Cambridge, UK.

KIM, J. 1993. *The nonreductivist's trouble with mental causation*. In "Supervenience and Mind", Cambridge University Press, Cambridge, pp. 336-57.

LETTVIN, J.Y. & MATURANA, H.R. & McCULLOCH, W.S. & PITTS, W.H. 1959. *What the frog's eye tells the frog's brain*. Proceedings of the Institute of Radio Engineers, pp. 1940-51.

LEWONTIN,1970. *The units of selection*. Annual Review of Ecology and Systematics, 1, p. 1-18.

MACDONALD, D. & PAPINEAU, D. 2006. *Teleosemantics*. Oxford University Press. Oxford.

MAMELI, M. 2004. *Nongenetic selection and nongenetic inheritance*. The British Journal for the Philosophy of Science, 55, 1, pp. 35–71.

MILLIKAN, R. G. 1984. *Language, thought and other biological categories*. MIT Press. Cambridge, MA.

—— 1989a. *In defense of proper functions*. Philosophy of Science. 56, 2, pp. 288-302.

—— 1989b. *Biosemantics.* The Journal of Philosophy, 86, pp. 281-97.

—— 1995. *Pushmi-Pullyu representations.* Philosophical Perspectives, 9, pp. 185-200.

—— 2001. *What has natural information to do with intentional representation?*. In "Evolution, naturalism and mind" (WALSH, ed.). Cambridge University Press, Cambridge, pp. 105-26.

—— 2004. *Varieties of reference.* MIT Press. Cambridge, MA.

—— 2007. *An Input condition for teleosemantics? Reply to Shea (and Godfrey-Smith).* Philosophy and Phenomeno1logical Research, LXXV, 2, pp. 436-55.

MOORE, G. E. 1903 [1993]. *Principia ethica*. Cambridge University Press. Cambridge.

NEANDER, K. 1991. *The teleological notion of function*. Australasian Journal of Philosophy, 69, pp. 458-64.

—— 1995. *Misrepresenting and malfunctioning.* Philosophical Studies, 79, pp. 109-41.

—— 2006. *Content for cognitive science.* In "Teleosemantics" (MACDONALD & PAPINEAU, eds.), Clarendon Press, Oxford, pp. 374-91.

—— 2012. *Teleological theories of mental content*. The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.), URL = < https://plato.stanford.edu/entries/content-teleological/>.

—— 2013. *Toward an informational teleosemantics.* In *Millikan and her critics* (RYDER & KINGSBURY & WILLIFORD, eds.), Wiley, Malden, MA, pp. 21-40.

—— 2017. *A Mark of the mental – in defense of informational teleosemantics*. MIT Press. Cambridge, MA.

O'KEEFE, J. & NADEL, L. 1978. *The hippocampus as a cognitive map*. Oxford University Press, Oxford.

O'KEEFE, J. & BURGESS, N. 2005. *Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells.* Hippocampus, 15, 7, pp. 853–866.

PALMER, S. 1978. *Fundamental aspects of cognitive representation.* In "Cognition and Categorization" (ROSCH & LLOYD, eds.), Lawrence Erlbaum, Hillsdale, NJ, pp. 259–303.

PAPINEAU, D. 1984. *Representation and explanation*. Philosophy of Science, 51, pp. 550-72.

—— 1993. *Philosophical naturalism*. Basil Blackwell. Oxford.

—— 1998. *Teleosemantics and indeterminacy.* Australasian Journal of Philosophy, 76, 1 pp. 1-14.

—— 2003. *Is representation rife?*. Ratio, XVI, p. 107-23.

—— 2006. *Naturalist theories of meaning.* In "The Oxford Handbook of Philosophy of Language" (LEPORE, E. & SMITH, B., eds), Oxford University Press, Oxford, pp. 175-188.

—— 2016. *Teleosemantics.* In "How biology shapes philosophy" (D. L. SMITH, org.), Cambridge University Press, Cambridge, pp. 95-120.

PRICE, C. 1998. *Determinate functions.* Noûs, 32, 1 pp. 54–75.

—— 2001. *Functions in mind: a theory of intentional content.* Oxford University Press. Oxford.

—— 2014. *Teleosemantics re-examined: content, explanation and norms.* Biology and Philosophy, 29, pp. 587-96.

PUTNAM, H. 1967. *Psychological predicates.* In "Art, Mind, and Religion" (W.H. CAPITAN & MERILL, eds.), Pittsburgh: University of Pittsburgh Press, pp. 37–48.

QUINE, W. V. O. 1960. *Word and object.* MIT Press. Cambridge, MA.

RAMSEY, F. 1927. *Facts and propositions*. In "Proceedings of the Aristotelian Society, the virtual issue" (LONGWORTH, G., ed.), No. 1, URL = <https://www.aristoteliansociety.org.uk/pdf/ramsey.pdf>.

RAMSEY, W. 2007. *Representation reconsidered*. Cambridge University Press. Cambridge.

RISTAU, C. 1991. *Aspects of cognitive ethology of an injury-feigning bird, the piping plover.* In "Cognitive ethology" (RISTAU, C., ed.), LEA Press, Hillsdale, pp. 91-126.

RAWLS, J. 1999. *A Theory of justice – revised edition*. Harvard University Press. Cambridge, MA.

RESCORLA, M. 2013. *Millikan on honeybee navigation and communication.* In *Millikan and her critics* (RYDER & KINGSBURY & WILLIFORD, eds.), Wiley, Malden, MA, pp. 87-102.

—— 2015. *The causal relevance of content to computation.* "Philosophy and Phenomenological Research", LXXXVIII, 1, pp. 173-208.

ROSENBERG, A. 2006. *Darwinian reductionism: or, how to stop worrying and love molecular biology.* University of Chicago Press. Chicago.

RYDER, D. 2009. *Problems of representation II: naturalizing content.* In "The Routledge Companion to the Philosophy of Psychology" (SYMONS & CALVO, eds.), Routledge, London, pp. 233-50.

SCHULTE, P. 2012. *How frogs see the world: putting Millikan's teleosemantics to the test*. Philosophia, 40, pp. 483–96

—— 2015. *Perceptual representations: a teleosemantic answer to the breadth-of-application problem.* Biological Philosophy, 30, pp. 119–36.

—— 2018. *Perceiving the world outside: how to solve the distality problem for informational teleosemantics.* Philosophical Quarterly, 68, 271, pp. 349–69.

SCHIFFER, S. 1982. *Intentional-based semantics.* Notre Dame Journal of Formal Logic, 23, 2, pp. 119-56.

SEARLE, J. 1983. *Intentionality.* Cambridge, UK: Cambridge University Press.

—— 1992. *The rediscovery of mind*. MIT Press. Cambridge, MA.

SELLARS, W. 1962. *Philosophy and the scientific image of man.* In "Frontiers of Science and Philosophy" (COLODNY, R., ed.), University of Pittsburgh Press, Pittsburgh, pp. 35–78.

SEYFARTH, R. et al. 1980. *Monkey responses to three different alarm calls: evidence of predator classification and semantic communicatio*n. Science, 210, 4471, pp. 801-03.

SOBER, E. 1984. *The nature of selection*. University of Chicago Press. Chicago.

—— 2008. *Fodor's bubbe meise against Darwinism.* Mind and Language, 23, 1, pp. 42-49.

—— 2010. *Natural selection, causality, and laws: what Fodor and Piatelli-Palmarini got wrong*. Philosophy of Science, 77, 4, pp. 594-607.

SHEA, N. 2007. *Consumers need information: supplementing teleosemantics with an input condition.* Philosophy and Phenomenological Research, LVXXV, 2, pp. 404-35.

—— 2013. *Naturalising representational content*. Philosophy Compass 8, 5, pp. 496-509.

SRINIVASAN, M. V. 2010. *Honey bees as a model for vision, perception and cognition.* Annual Review of Entomology, 55, pp. 267-84.

STERELNY, K. 1990. *The Representational theory of mind*. Basil Blackwell. Cambridge, MA.

—— 1995. *Basic minds*. Philosophical Perspectives, 9, pp. 251-70.

STERELNY, K. & GRIFFITHS, P. E. 1999. *Sex and death: an introduction to philosophy of biology*. The University of Chicago Press. Chicago and London.

STICH, S. P. 1983. *From folk psychology to cognitive science*. MIT Press. Cambridge, MA.

STICH, S. P. & LAURENCE, S. 1994. *Intentionality and naturalism*. Midwest Studies in Philosophy, XIX, 1, pp. 159-82.

STRAWSON, G. 1986. *Freedom and belief*. Oxford University Press. Oxford.

VICENTE, A. 2012. *Burge on representation and biological function*. Thought, 1, 2, pp. 125-33.

von FRISCH, K. 1967. *The dance language and orientation of the bees*. Harvard University Press. Cambridge, MA.

von ECKARDT, B. 1993. *What is cognitive science?*. MIT Press. Cambridge, MA.

WHYTE, J. 1990. *Success semantics*. Analysis. 50, 3, pp. 149-57.

WRIGHT, L. 1973. *Functions*. The Philosophical Review, 82, pp. 139-68.